Quantile Regression in Risk Calibration

DISSERTATION

zur Erlangung des akademischen Grades
doctor rerum politicarum
(Doktor der Wirtschaftswissenschaft)

eingereicht an der
Wirtschaftswissenschaftlichen Fakultät
der Humboldt-Universität zu Berlin

von
M.B.A. Shih-Kang Chao

Präsident der Humboldt-Universität zu Berlin:
Prof. Dr. Jan-Hendrik Olbertz

Dekan der Wirtschaftswissenschaftlichen Fakultät:
Prof. Dr. Ulrich Kamecke

Gutachter:

1. Prof. Dr. Wolfgang Karl Härdle
2. Prof. Dr. Vladimir Spokoiny

Tag des Kolloquiums: 2 Juni, 2015

Abstract

Quantile regression studies the *conditional quantile function* $Q_{Y|X}(\tau)$ on X at level τ which satisfies $F_{Y|X}\left[Q_{Y|X}(\tau)\right] = \tau$, where $F_{Y|X}$ is the conditional CDF of Y given X, $\forall \tau \in (0,1)$. Quantile regression allows for a closer inspection of the conditional distribution beyond the conditional moments. This technique is particularly useful in, for example, the Value-at-Risk (VaR) which the Basel accords (2011) require all banks to report, or the "quantile treatment effect" and "conditional stochastic dominance (CSD)" which are economic concepts in measuring the effectiveness of a government policy or a medical treatment.

Given its value of applicability, to develop the technique of quantile regression is, however, more challenging than mean regression. It is necessary to be adept with general regression problems and M-estimators; additionally one needs to deal with non-smooth loss functions. In this dissertation, chapter 2 is devoted to empirical risk management during financial crises using quantile regression. Chapter 3 and 4 address the issue of high-dimensionality and the nonparametric technique of quantile regression.

Chapter 2 applies nonparametric confidence bands for quantile functions to investigate the tail dependence of stock returns. It is shown that strong *nonlinear* correlation exists when stock prices drop, confirming the fact that in financial crises, firms are more dependent on each other than when the market is booming. This sheds light on the risk management of counterparty risk.

In Chapter 3, motivated by applications in economics like quantile treatment effects, or conditional stochastic dominance, we focus on the construction of confidence corridors for nonparametric *multivariate* kernel quantile and expectile regression functions. Through an uniform kernel Bahadur representation for M-estimators, strong Gaussian approximation and asymptotic extreme value theory we derive the asymptotic confidence corridor for the nonparametric kernel conditional quantile/expectile functions. We find that the bands for quantile/expectile functions are wide when τ is close to 0 and 1 due to the variance of the estimator. The coverage ratios given by the asymptotic confidence corridors are meager. To deal with this issue, we propose a novel smoothing bootstrap which gives satisfactory coverage ratios while keeping the size of the confidence corridors in a reasonable range. Our method contributes to the differentiation between the "risk reduction CSD" and "potential enhancement CSD", which is not possible by using techniques based on previous research in CSD like Delgado and Escanciano (2013). This differentiation is crucial as the two types of CSD may induce different utility to the government and citizens. After applying our method to the data set from National Supported Work Demonstration, a temporary internship program offered to disadvantaged workers, it is found that this program tends to be "potential enhancement CSD" and it may not help foster the employment of less capable people as much as get the more capable people higher pay.

Chapter 4 deals with factorisable multivariate quantile regression model. Factor models appear frequently in a variety of fields in science. In economics, the Capital

Asset Pricing Model (CAPM) and Arbitrage Pricing Theory (APT) are famous examples. When the factors are not identified ex-ante, reduced-rank multivariate regression, in which the response variables and input variables are both vectors linked by a matrix, can be applied to estimate the factors and the model. Yuan et al. (2007), Negahban and Wainwright (2011) and Bunea et al. (2011) show that using nuclear norm or rank regularization the number of factors can be estimated with high probability. However, the models studied so far only focus on conditional expected values and give little information for the conditional distributions. For $\tau \in (0, 1)$, the conditional τ-quantile functions, particularly for τ close to 0 or 1, are crucial in many applications, such as risk management or weather analysis. In this chapter, the estimation of large multivariate quantile regression models regularized by nuclear norm is considered. The rank of the coefficient matrix is interpreted as the factors for the tail event functions, and is sparse in the spirit of CAPM and APT. Hence, we call the estimated quantile functions FASTEC: FActorisable Sparse Tail Event Curves. Our approach can be viewed as a multi-task learning problem for quantile regression which gives more accurate estimations of quantiles than single-task learning by incorporating information from other variables. Moreover, our approach allows for summarizing the behavior of a group of variables into 'factors'. Our technique can also be easily extended to nonparametric multivariate quantile estimation through the use of sieve method.

As the empirical loss function and the nuclear norm are both non-smooth, an efficient algorithm for estimation, which combines smoothing techniques and effective proximal gradient methods, is developed, for which explicit deterministic convergence rates are derived. It is shown that the estimator has nonasymptotic oracle properties under rank sparsity condition. The technique is applied to a multivariate variation of the famous Conditional Autoregressive Value-at-Risk (CAViaR) model of Engle and Manganelli (2004), which is called Sparse Asymmetric Conditional Value-at-Risk (SAMCVaR). With data consisting of stock prices of global financial firms from 2007 to 2010, our method is able to identify the major risk contributors and market sensitive firms. We also apply the nonparametric multivariate quantile regression to analyze the nationwide Chinese temperature in 2008 and classify the patterns of seasonality of the demeaned temperature time series.

Zusammenfassung

Die Quantilsregression untersucht die Quantilfunktion $Q_{Y|X}(\tau)$, sodass $\forall\tau \in (0,1)$, $F_{Y|X}[Q_{Y|X}(\tau)] = \tau$ erfüllt ist, wobei $F_{Y|X}$ die bedingte Verteilungsfunktion von Y gegeben X ist. Die Quantilsregression ermöglicht eine genauere Betrachtung der bedingten Verteilung über die bedingten Momente hinaus. Diese Technik ist in vielerlei Hinsicht nützlich: beispielsweise für das Risikomaß *Value-at-Risk (VaR)*, welches nach dem Basler Akkord (2011) von allen Banken angegeben werden muss, für "Quantil treatment-effects" und die "bedingte stochastische Dominanz (CSD)", welches wirtschaftliche Konzepte zur Messung der Effektivität einer Regierungspolitik oder einer medizinischen Behandlung sind.

Die Entwicklung eines Verfahrens zur Quantilsregression stellt jedoch eine größere Herausforderung dar, als die Regression zur Mitte. Allgemeine Regressionsprobleme und M-Schätzer erfordern einen versierten Umgang und es muss sich mit nichtglatten Verlustfunktionen beschäftigt werden. Kapitel 2 behandelt den Einsatz der Quantilsregression im empirischen Risikomanagement während einer Finanzkrise. Kapitel 3 und 4 befassen sich mit dem Problem der höheren Dimensionalität und nichtparametrischen Techniken der Quantilsregression.

In Kapitel 2 werden nichtparametrische Konfidenzbereiche für Quantilfunktionen angewendet, um die Abhängigkeit von Aktienrenditen in den Rändern der Verteilung zu untersuchen. Es wird gezeigt, dass eine starke nichtlineare Korrelation besteht, wenn Aktienkurse fallen. Dies ist im Einklang mir der Tatsache, dass Firmen in Finanzkrisen stärker voneinander abhängig sind, als wenn der Markt boomt und gibt Aufschluss über das Risikomanagement von Kontrahentenrisiko.

Kapitel 3 konzentriert sich auf die Herleitung von Konfidenzbereichen für nichtparametrische, multivariate Kernel-Quantile und Expektilregressionsfunktionen, motiviert durch Anwendungen, wie dem Quantil treatment-effect oder der bedingten stochastischen Dominanz. Mit Hilfe einer *Uniform Kernel Bahadur Representation* für M-Schätzer, *Strong Gaussian Approximation* und der asymptotischen Extremwerttheorie leiten wir den asymptotischen Konfidenzbereich für nicht parametrische kernelbedingte Quantil-/ Expektilfunktionen her. Es zeigt sich, dass die Bereiche für die Quantil- und Expektilfunktionen groß sind, wenn τ aufgrund der Varianz des Schätzers nahe bei 0 oder 1 liegt. Die *Coverage Ratios* der asymptotischen Konfidenzbereiche sind gering. Um dieses Problem anzugehen, schlagen wir eine neue Bootstrap-Glättung vor, die zufriedenstellende Coverage Ratios liefert, während die Größe der Konfidenzbereiche in einem angemessenen Bereich bleibt. Unsere Methode trägt zur Differenzierung zwischen "Risk Reduction CSD" und "Potential Enhancement CSD" bei, was mit Techniken früherer Forschungen zu CSD, wie der von Delgado and Escanciano (2013), nicht möglich ist. Diese Unterscheidung ist wichtig, da die beiden Arten von CSD unterschiedlichen Nutzen für Staat und Einwohner herbeiführen. Nach Anwendung unserer Methode auf den Datensatz der National Supported Work Demonstration aus den 1970er Jahren, stellt man fest, dass das Programm eher eine "potenziell verbesserte CSD" ist und es nicht unbedingt dazu beiträgt, die Beschäftigung von gering leistungsfähigen Menschen zu fördern.

Kapitel 4 befasst sich mit faktorisierbaren multivariaten Modellen der Quantilsregression. Faktormodelle werden häufig in einer Vielzahl von Wissenschaftsfeldern verwendet. Beispiele aus der Wirtschaft sind etwa das Capital Asset Pricing Model (CAPM) und Arbitrage Pricing Theory (APT). Wenn die Faktoren unbekannt sind, kann zur Schätzung und Bestimmung des Modells die reduced-rank multivariate Regression angewendet werden, bei der sowohl die Ziel-, als auch die Eingangsgrößen über eine Matrix gekoppelte Vektoren sind. Yuan et al. (2007), Negahban and Wainwright (2011) and Bunea et al. (2011) zeigen, dass unter Einsatz der Ky-Fan-Norm oder Rang Regularisierung die Anzahl der Faktoren mit hoher Wahrscheinlichkeit geschätzt werden kann. Allerdings konzentrieren sich die bisher untersuchten Modelle nur auf bedingte Erwartungswerte und geben wenig Informationen über die bedingten Verteilungen. Für $\tau \in (0, 1)$ sind die bedingten τ Quantil-Funktionen, insbesondere für τ nahe 0 oder 1, für viele Anwendungen von entscheidender Bedeutung, wie z.B. für das Risikomanagement oder die Wetteranalyse. In der vorliegenden Studie wird die Schätzung von großen multivariaten über die Ky-Fan-Norm regularisierten Quantilsregressionsmodellen betrachtet. Der Rang der Koeffizientenmatrix wird als die Faktoren für die Randereignisfunktionen interpretiert und ist sparse im Sinne des CAPM und APT. Daher nennen wir die geschätzten Quantilfunktionen *FASTEC: FActorisable Sparse Tail Event Curves*. Unsere Methode kann als ein Multi-Task-Lernproblem für Quantilsregression betrachtet werden, welches durch die Einbeziehung von Informationen aus anderen Variablen eine genauere Schätzung liefert als beim Single-Task-Lernen. Darüber hinaus ermöglicht unser Ansatz die Zusammenfassung des Verhaltens einer Gruppe von Variablen durch "Faktoren". Unsere Technik kann für nicht-parametrische, multivariate Qunatilschätzungen durch die Anwendung der Sieb-Methode einfach erweitert werden.

Da die empirische Verlustfunktion und die Ky-Fan-Norm beide nicht glatt sind, wird in diesem Kapitel ein effizienter Schätzungsalgorithmus entwickelt, der generale Glättungstechniken und effektive proximale Gradientenverfahren kombiniert. Daraus werden dann explizite deterministische Konvergenzraten abgeleitet. Es wird gezeigt, dass der Schätzer nicht asymptotische *oracle Properties* mit *Rank Sparsity Condition* aufweist. Die Technik wird auf eine multivariate Variante des bekannten *Conditional Autoregressive Value-at-Risk (CAViaR)* Modells von Engle and Manganelli (2004) angewendet, welches *Sparse Asymmetric Conditional Value-at-Risk (SAMCVaR)* genannt wird. Mit einem Datensatz, bestehend aus Aktienkursen globaler Finanzunternehmen von 2007 bis 2010, werden mit unserer Methode Marktrisikofaktoren und marktsensitive Unternehmen identifiziert. Wir wenden außerdem die nicht-parametrische, multivariate Quantilregression an, um die landesweite Temperatur im Jahr 2008 in China zu analysieren und Saisonmuster der mittelwertbereinigten Temperaturzeitreihe zu klassifizieren.

Acknowledgment

I cannot have finished this dissertation without the support from my colleagues and family. There are a few persons who I owe great amount of debt of gratitude.

Professor Dr. Wolfgang Karl Härdle has been a great mentor. I thank him for the generous financial support and the share of knowledge over the past years. He shaped my view of the world of statistics, and nurtured my taste for good research. I am also greatly influenced by the works and ideas of my second supervisor Prof. Dr. Spokoiny. In addition, I thank the kind host and share of ideas from Dr. Katharina Proksch and Prof. Dr. Holger Dette when I visited Ruhr-Universitt Bochum. I learned a great deal from their rigorous and elegant style for developing statistical theory. I would also like to thank Prof. Ming Yuan from the University of Wisconsin-Madison, who has been a great guide showing me the key ideas of high-dimensional statistics.

I would also like to extend my thanks to Prof. Dr. Markus Reiß. I benefited a lot from his wonderful course "Nichtparametrische Statistik" (Nonparametric statistics) held in the winter semester of 2012. His dedication as a teacher will be my guide when I become a teacher myself.

My colleagues in Ladislaus von Bortkiewicz Chair of Statistics and CRC 649 "Economic Risk", Humboldt-Universität zu Berlin have guided and helped me through my years as a Ph.D. student. Particularly, Dr. Andrija Mihoci has always been helpful when I am in need, I would like to thank him for his patience and time. I thank Prof. Dr. Ostap Okhrin for giving me many advices when I instructed courses with him. The financial support of CRC 649 "Economic Risk", Humboldt-Universität zu Berlin is also gratefully acknowledged.

The financial support from the Berlin Doctoral Program for Economics and Management Science is gratefully acknowledged. I would also like to thank my comrades: Simon Jurkatis, Tsung-Hsien Lee, Lei Fang. I benefited a lot from their friendship. Hopefully we all have a career that we enjoy.

Last but not the least, I thank my parents, who have always been supportive in my life. I also thank my wife Limei for her company and love.

Contents

1 Introduction **1**

2 Quantile Regression in Risk Calibration **5**
2.1 Introduction . . . 5
2.2 Methodology . . . 9
2.2.1 Constructing Partial Linear Model (PLM) for CoVaR . . . 9
2.2.2 Backtesting . . . 13
2.2.3 Risk contribution measure . . . 15
2.3 Results . . . 16
2.3.1 CoVaR estimation . . . 16
2.3.2 Backtesting . . . 19
2.3.3 Global risk contribution . . . 21
2.4 Conclusion . . . 23

3 Confidence Corridors for Generalized Quantile Regression **25**
3.1 Introduction . . . 25
3.2 Asymptotic confidence corridors . . . 27
3.2.1 Prerequisites . . . 27
3.2.2 Asymptotic results . . . 28
3.2.3 Estimating the scaling factors . . . 32
3.3 Bootstrap confidence corridors . . . 34
3.3.1 Asymptotic theory . . . 34
3.3.2 Implementation . . . 37
3.4 A simulation study . . . 38
3.5 Application: a treatment effect study . . . 43

4 FASTEC: Factorisable Sparse Tail Event Curves **53**
4.1 Introduction . . . 53
4.1.1 Related work . . . 57
4.1.2 Notations of this chapter . . . 59
4.2 Factorizable sparse multivariate quantile regression . . . 59
4.3 Estimation . . . 61
4.4 Oracle inequalities . . . 65
4.5 Tuning . . . 71
4.6 Simulation . . . 72

4.6.1 Symmetric models . . . 72
4.6.2 Asymmetric models . . . 74
4.7 Real data application: SAMCVaR model . . . 81
4.7.1 Model . . . 81
4.7.2 Data and tuning . . . 82
4.7.3 Results . . . 84
4.8 Factor curve model . . . 92
4.8.1 Model . . . 92
4.8.2 Estimation . . . 93
4.8.3 Application: Chinese temperature data . . . 93

Bibliography **106**

A Supplementary materials for Chapter 2 **107**
A.1 Locally Linear Quantile Regression (LLQR) . . . 107
A.2 Confidence band for nonparametric quantile estimator . . . 109
A.3 PLM model estimation . . . 110

B Supplementary materials for Chapter 3 **111**
B.1 Proof of Theorems . . . 111
B.1.1 Proof of Theorem 3.2.1 . . . 114
B.1.2 Proof of Theorem 3.2.4 . . . 123
B.1.3 Proof of Lemma 3.2.8 . . . 127
B.1.4 Proof of Theorem 3.3.1 . . . 130
B.2 Supporting lemmas . . . 138

C Supplementary materials for Chapter 4 **141**
C.1 Proof for algorithmic convergence analysis . . . 141
C.1.1 Proof of Theorem 4.3.2 . . . 141
C.1.2 Proof of Theorem 4.3.3 . . . 141
C.2 Proof of oracle inequalities . . . 142
C.2.1 Proof of Lemma 4.4.1 . . . 142
C.2.2 Proof of Lemma 4.4.2 . . . 143
C.2.3 Proof of Lemma 4.4.3 . . . 143
C.2.4 Proof of Lemma 4.4.5 . . . 145
C.3 Supplementary lemmas . . . 146

List of Figures

2.1.1 Goldman Sachs (GS) and Citigroup (C) weekly returns 0.05(left) and 0.1(right) quantile functions. The y-axis is GS daily returns and the x-axis is the C daily returns. The blue curve are the locally linear quantile regression curves (see Appendix A.1). The locally linear quantile regression bandwidth are 0.1026 and 0.0942. The red lines are the linear parametric quantile regression line. The antique white dashed curves are the asymptotic confidence band (see Section A.2) with significance level 0.05. The sample size $N = 546$. 8

2.2.1 The scatter plots of GS daily returns to the 7 market variables with the LLQR curves. The bandwidths are selected by the method described in Appendix A.1. The LLQR bandwidths are 0.1101, 0.1668, 0.2449, 0.0053, 0.0088, 0.0295 and 0.0569. The data period is from August 4, 2006 to August 4, 2011. $N = 1260$. $\tau = 0.05$ 11

2.2.2 (Continued from Figure 2.2.1) 12

2.2.3 The nonparametric part $\hat{l}_{GS|C}(\cdot)$ of the PLM estimation. The y-axis is the GS daily returns. The x-axis is the C daily returns. The blue curve is the LLQR quantile curve. The red line is the linear parametric quantile line. The magenta dashed curves are the asymptotic confidence band with significance level 0.05. The data is from June 25, 2008 to December 23, 2009. 378 observations. Bandwidth =0.1255. $\tau = 0.05$. 13

2.3.1 The $\widehat{VaR}_{GS,t}$. The red line is the $\widehat{VaR}_{GS,t}$ and blue stars are daily returns of GS. The dark green curve is the meadian smoother of the $\widehat{VaR}_{GS,t}$ curve with h=2.75. $\tau = 0.05$. The window size is 252 days. . 17

2.3.2 The CoVaR of GS given the VaR of C. The gray dots are daily returns of GS. The light green dashed curve is the $\widehat{CoVaR}^{PLM}_{GS|C,t}$. The blue curve is the median LLQR smoother of the light green dashed curve with $h = 3.19$. The cyan dashed curve is the $\widehat{CoVaR}^{AB}_{GS|C,t}$. The purple curve is the median LLQR smoother of the cyan dashed curve with $h = 3.90$. The red curve is the $\widehat{VaR}_{GS,t}$. $\tau = 0.05$. The moving window size is 126 days. 18

2.3.3 LLQR bandwidth in the moving daily estimation of $\widehat{CoVaR}^{PLM}_{GS|C,t}$. The average bandwidth is 0.24. 19

2.3.4 The timings of violations $\{t : I_t = 1\}$. The top circles are the violations of the $\widehat{CoVaR}^{PLM}_{GS|C,t}$, totally 95 violations. The middle squares are the violations of $\widehat{CoVaR}^{AB}_{GS|C,t}$, totally 98 violations. The bottom stars are the violations of $\widehat{VaR}_{GS,t}$, totally 109 violations. Overall data $N = 1260$. 20
2.3.5 The timings of violations $\{t : I_t = 1\}$. The top circles are the violations of $\widehat{CoVaR}^{PLM}_{GS|SP,t}$, totally 123 violations. The middle squares are the violations of $\widehat{CoVaR}^{AB}_{GS|SP,t}$, totally 39 violations. The bottom stars are the violations of $\widehat{VaR}_{GS,t}$, totally 109 violations. Overall data $N = 1260$. 20
2.3.6 The $MCR^{\tau_1}_j$, $\tau = 0.5$. j:CAC, FTSE, DAX, Heng Seng, S&P500 and NIKKEI225. The global market return is approximated by MSCI World. 22
2.3.7 The $MCR^{\tau_2}_j$, $\tau = 0.05$. j:CAC, FTSE, DAX, Heng Seng, S&P500 and NIKKEI225. The global market return is approximated by MSCI World. 23

3.5.1 The illustrations for the two possible types of stochastic dominance. In the left figure, the 0.1 quantile improves (downside risk reduction) more dramatically than the 0.9 quantile (upside potential increase), as the distance between A and A' is greater than that between B and B'. For the right picture the interpretation is just the opposite. . . . 44
3.5.2 Unconditional empirical density function (left) and distribution function (right) of the difference of earnings from 1975 to 1978. The dashed line is associated with the control group and the solid line is associated with the treatment group. 45
3.5.3 Nonparametric quantile regression estimates and CCs for the changes in earnings between 1975-1978 as a function of age. The solid dark lines correspond to the conditional quantile of the treatment group and the solid light lines sandwich its CC, and the dashed dark lines correspond to the conditional quantiles of the control group and the solid light lines sandwich its CC. 48
3.5.4 Nonparametric quantile regression estimates and CCs for the changes in earnings between 1975-1978 as a function of years of schooling. The solid dark lines correspond to the conditional quantile of the treatment group and the solid light lines sandwich its CC, and the dashed dark lines correspond to the conditional quantiles of the control group and the solid light lines sandwich its CC. 49
3.5.5 The CCs for the treatment group and the control group. The net surface corresponds to the control group quantile CC and the solid surface corresponds to the treatment group quantile CC. 50
3.5.6 The conditional quantiles (solid surfaces) for the treatment group and the CCs (net surfaces) for the control group. 51

4.1.1 The variable simulated by (4.1.1). The left is Y_1 bounded above by 0 and the left is Y_{101} bounded below by 0. 54
4.1.2 The PCA biplot on data $\mathbf{Y}$. PCA is based on the covariance and does not capture the pattern in the quantiles of the distribution. 55
4.1.3 The first factor of 1% (black) and 99% (blue) quantiles of data $\mathbf{Y}$(left) and the factor loadings(right). Variables have close distance on the right figure have similar change in τ-range, $\tau = 1\%$. 56
4.3.1 The solid line is the function $\psi_\tau(u) = \tau - \mathbf{1}(u \leq 0)$ with $\tau = 0.5$, which has a jump at the origin. The dashed line corresponding to the smoothing gradient $[[\kappa^{-1}(\mathbf{Y} - \mathbf{X}\boldsymbol{\Gamma})]]_\tau$ associated with $\kappa = 0.5$. As κ decreases to 0.05, we observe that the smoothing approximation function is closer to $\psi_\tau(u)$. 64
4.6.1 The plot of all 500 marginal densities of $\boldsymbol{Y}_i$ in asymmetric models. The left figure is associated with Model AMS in which the densities tend to be asymmetric (thick right tails and thin left tails). The right figure is associated with Model AES in which the densities are more symmetric. 75
4.6.2 The symmetric Model LS. The horizontal axis is τ. The true number of factors is 125. 76
4.6.3 The symmetric Model MS. The horizontal axis is τ. The true number of factors is 10. 77
4.6.4 The symmetric Model ES. The horizontal axis is τ. The true number of factors is 1. 78
4.6.5 The asymmetric Model AES. The horizontal axis is τ. The true number of factors is 2 for $\tau < 0.5$ and 10 for $\tau > 0.5$. 0 for $\tau = 0.5$. . 79
4.6.6 The asymmetric Model AMS. The horizontal axis is τ. The true number of factors is 2 for $\tau < 0.5$ and 10 for $\tau > 0.5$. 0 for $\tau = 0.5$. . 80
4.7.1 The upper figure shows the time series plots of the 230 global financial institutions with different grey level distributions and thicknesses. The lower figure shows the time series of VIX. 83
4.7.2 The time series plots for the first 2 factors. The black lines corresponds to 1% quantile factors and the blue lines corresponds to 99% quantile factors. 84
4.7.3 The magnitude of contribution to the first factor of 1% and 99% MQR from the 230+230 covariates. The firm name and the black dots denote the squared log return $Y^2_{t-1,j}$. Red dots and firm name with "–" denote the lag negative return $Y^-_{t-1,j}$. 85
4.7.4 The factor loadings of 230 firms on the first factors $\boldsymbol{f}_1(0.01)$ and $\boldsymbol{f}_1(0.99)$. 86
4.7.5 The magnitude of contribution to the second factor of 1% and 99% MQR from the 230+230 covariates. The firm name and the black dots denote the squared log return $Y^2_{t-1,j}$. Red dots and firm name with "–" denote the lag negative return $Y^-_{t-1,j}$. 87

4.7.6 The factor loadings of 230 firms on the second factors $\boldsymbol{f}_2(0.01)$ and $\boldsymbol{f}_2(0.99)$. 87
4.7.7 The magnitude of contribution to the *first* and *second* factor of 1% MQR from the 230+230 covariates. The firm name and the black dots denote the squared log return $Y^2_{t-1,j}$. Red dots and firm name with "−" denote the lag negative return $Y^-_{t-1,j}$. 88
4.7.8 The factor loadings of 230 firms on the second factors $\boldsymbol{f}_1(0.01)$ and $\boldsymbol{f}_2(0.01)$ of 1% MQR. 89
4.7.9 Plots of individual asset time series and their 1% and 99% fitted quantiles. 90
4.7.10 Plots of individual asset time series and their 1% and 99% fitted quantiles (continued). 91
4.8.1 The temperature time series in excess to national mean of the 159 weather stations around China with different grey level distributions and thicknesses and the temperature trend curve. 94
4.8.2 The time series plots for the first 4 factors. The black lines corresponds to 1% quantile factors and the blue lines corresponds to 99% quantile factors. 95
4.8.3 The plot of weather stations based on their factor loadings to 1% and 99% multivariate quantile regression. Each point denotes a weather station somewhere in China. 96
4.8.4 Plots of temperature observations, 1%, and 99% temperature quantile curves of the three weather stations in the year 2008. The location of the weather stations are marked in the upper left map of China. . 97

A.1.1 This figure presents the check function. The dotted line is u^2. The dashed and solid lines are check functions $\rho_\tau(u)$ with $\tau = 0.5$ and 0.9 respectively. 108
A.1.2 GS and C weekly returns 0.90(left) and 0.95(right) quantile functions. The y-axis is GS daily returns and the x-axis is the C daily returns. The blue curves are the LLQR curves (see Appendix A.1). The LLQR bandwidths are 0.0942 and 0.1026. The red lines are the linear parametric quantile regression line. The antique white curves are the asymptotic confidence band (see Appendix A.2) with significance level 0.05. $n = 546$. 109

List of Tables

2.3.1 VaR/CoVaR summary statistics. The overall period is from August 4, 2006 to August 4, 2011. The crisis period is from August 4, 2008 to August 4, 2009. The numbers in the table are scaled up by 10^2. . . 18
2.3.2 Goldman Sachs VaR/CoVaR backtesting p-values. The overall period is from August 4, 2006 to August 4, 2011. The crisis period is from August 4, 2008 to August 4, 2009. LB(1) and LB(5) are the Ljung-Box tests of lags 1 and 5. L(1) and L(5) are the Lobato tests of lags 1 and 5. CaViaR-overall and CaViaR-crisis are two CaViaR tests described in Section 2.2.2 applied on the two data periods. 21

3.4.1 Nonparametric quantile model coverage probabilities. The nominal coverage is 95%. The number in the parentheses is the volume of the confidence corridor. The asymptotic method corresponds to the asymptotic quantile regression CC and bootstrap method corresponds to quantile regression bootstrap CC. 39
3.4.2 Nonparametric expectile model coverage probability. The nominal coverage is 95%. The number in the parentheses is the volume of the confidence corridor. The asymptotic method corresponds to the asymptotic expectile regression CC and bootstrap method corresponds to expectile regression bootstrap CC. 40
3.4.3 Proportion in 2000 iteration that the coverage of $\geq 95\%$ grid points for nonparametric mean model, using the bootstrap method of Hall and Horowitz (2013). The nominal coverage is 95%. The number in the parentheses is the volume of the confidence corridor. 42
3.5.1 The unconditional sample quantiles of treatment and control groups. 46
3.5.2 The two sample empirical cdf tests results for treatment and control groups. 46

4.7.1 Summary of firm characteristics. There are three geographical categories: Europe, North America and Asia, and also three industrial categories: bank, financial service and insurance. 83

Chapter 1

Introduction

Quantile regression studies the *conditional quantile function* $Q_{Y|X}(\tau)$ on X at level τ which satisfies $F_{Y|X}\left[Q_{Y|X}(\tau)\right] = \tau$, where $F_{Y|X}$ is the conditional CDF of Y given X, $\forall \tau \in (0,1)$. In comparison to usual regression analysis, quantile regression allows for a closer inspection of the conditional distribution. This technique is particularly useful in, for example, the Value-at-Risk (VaR) which the Basel accords (2011) require banks to report. VaR is defined as the τ-quantile of the return distribution at time $t+d$ conditioned on the information set $\mathcal{F}_t$:

$$VaR_{t+d}^{\tau} \stackrel{\text{def}}{=} \inf\{x \in \mathbb{R} : \mathrm{P}(X_{t+d} \leq x|\mathcal{F}_t) \geq \tau\}, \text{ for } 0 < \tau < 1,$$

where X_t is the asset return and $\mathcal{F}_t$ is the information set at time t.

In econometrics, quantile regression are useful for studying the "quantile treatment effect" and "conditional stochastic dominance (CSD)". To see the relation of quantile regression to the quantile treatment effect, Lehmann (1975) proposed a general model for modeling the treatment response. Suppose the treatment adds $\Delta(y)$ to the treatment group, the distribution function $F_1(y)$ of the group being treated is related to the distribution function $F_0(y)$ of the control (untreated) group by $F_1(y) = F_0\{y+\Delta(y)\}$. Doksum (1974) shows that if setting $\tau = F_1(y)$ for $0 < \tau < 1$, applying F_0^{-1} to the both sides of $F_1(y) = F_0\{y+\Delta(y)\}$ gives $\Delta_\tau = F_0^{-1}(\tau) - F_1^{-1}(\tau)$, which is the quantile treatment effect. If we control for the covariates $\boldsymbol{X}$, we have the *conditional quantile treatment effect* $\Delta_\tau(\boldsymbol{x}) = F_{0|\boldsymbol{X}}^{-1}(\tau|\boldsymbol{x}) - F_{1|\boldsymbol{X}}^{-1}(\tau|\boldsymbol{x})$. Quantile regression can be applied to estimate the conditional quantiles $F_{0|\boldsymbol{X}}^{-1}(\tau|\boldsymbol{x})$ and $F_{1|\boldsymbol{X}}^{-1}(\tau|\boldsymbol{x})$ using the data of control and treatment groups, and $\Delta_\tau(\boldsymbol{x})$ can be estimated.

The concept of conditional stochastic dominance can be viewed as an extension of the conditional quantile treatment effect. According to Delgado and Escanciano (2013), Y_1 conditionally stochastically dominates Y_0 if $F_{1|\boldsymbol{X}}(y|\boldsymbol{x}) \leq F_{0|\boldsymbol{X}}(y|\boldsymbol{x})$ a.s. for all $y, \boldsymbol{x}$, where $\boldsymbol{x} \in \mathbb{R}^d$. Take $\tau = F_{0|\boldsymbol{X}}^{-1}(y|\boldsymbol{x})$, applying $F_{1|\boldsymbol{X}}^{-1}$ to the both sides of $F_{1|\boldsymbol{X}}(y|\boldsymbol{x}) \leq F_{0|\boldsymbol{X}}(y|\boldsymbol{x})$ yields the equivalent definition for condition stochastic dominance on $\boldsymbol{x}$: $F_{0|\boldsymbol{X}}^{-1}(\tau|\boldsymbol{x}) \leq F_{1|\boldsymbol{X}}^{-1}(\tau|\boldsymbol{x})$, the conditional $\tau-$quantile of Y_0 is less than that of Y_1, for all $\tau, \boldsymbol{x}$ a.s. This again can be estimated via quantile regression.

Chapter 2 applies nonparametric confidence bands for quantile functions to investigate the tail dependence of stock returns. Our idea is motivated by the CoVaR of Adrian and Brunnermeier (2011), in which a two-step quantile regression model is proposed to measure the systemic risk. However, we show that strong *nonlinear* correlation exists when stock prices drop. This also confirm the fact that in financial crises, firms are more dependent on each other than when the market is booming, but we do not observe much dependence when the market booms. We also show with daily stock returns of large market participants that the VaR incorporating the nonlinear dependence captures the risk during financial crisis. This sheds light on managing the counterparty risk.

To measure the quantile treatment effects and conditional stochastic dominance one needs statistical techniques to test whether the two conditional quantiles are the same. In Chapter 3, we focus on the construction of confidence corridors for nonparametric *multivariate* kernel quantile and expectile regression functions. Simultaneous confidence bands for nonparametric estimators have been constructed for many model settings. For example, Claeskens and Van Keilegom (2003) proposed the uniform confidence bands for mean regression curves and their derivatives. In time series setting, Liu and Wu (2010) constructed the uniform confidence bands for nonparametric density and mean estimator. In this chapter, through an uniform kernel Bahadur representation for M-estimators, strong Gaussian approximation and asymptotic extreme value theory, we derive the asymptotic confidence corridor for the nonparametric kernel conditional quantile/expectile functions. We find that the bands for quantile/expectile functions are wide when τ is close to 0 and 1 due to the variance of the estimator. The coverage ratios given by the asymptotic confidence corridors are meager, and the coverage ratios of usual nonparametric bootstrap for quantile regression estimator also perform poorly. To deal with this issue, we propose a novel smoothing bootstrap which gives satisfactory coverage ratios while keeping the size of the confidence corridors in a reasonable range. Our method contributes to the differentiation between the "risk reduction CSD" and "potential enhancement CSD", which is not possible by using techniques based on previous research in CSD like Delgado and Escanciano (2013). This differentiation is crucial as the two types of CSD may induce different utility to the government and citizens. After applying our method to the data set from National Supported Work Demonstration, a temporary internship program offered to disadvantaged workers, it is found that this program tends to be "potential enhancement CSD" and it may not help foster the employment of less capable people as much as get the more capable people higher pay.

In Chapter 4, we deal with high-dimensional multivariate quantile analysis. High-dimensional multivariate quantile analysis is crucial for many applications, such as risk management and weather analysis. In these applications, quantile functions $q_Y(\tau)$ of random variable Y such that $\mathrm{P}\{Y \leq q_Y(\tau)\} = \tau$ at the "tail" of the distribution, namely at τ close 0 or 1, such as $\tau = 1\%, 5\%$ or $\tau = 95\%, 99\%$, is of great interest. The quantile at level τ can be interpreted as the lower (upper) bound with confidence level $1 - \tau$ (τ) of the possible outcome of a random variable, and

the difference of $(q_Y(\tau), q_Y(1-\tau))$ can be interpreted as τ-range, with $\tau = 25\%$ being the special case of interquartile range. While covariance based methods such as principal component analysis do not yield information for the bounds, and are easily corrupted if data are highly skewed and present outliers. We propose a conditional quantile based method which enables *localized* analysis on quantiles and *global* comovement analysis for τ-range for high-dimensional data with factors. We call our method FASTEC: FActorisable Sparse Tail Event Curves.

The technique is implemented by factorising the multivariate quantile regression with nuclear norm regularization. As the empirical loss function and the nuclear norm are non-smooth, an efficient algorithm which combines smoothing techniques and effective proximal gradient methods is developed, for which explicit deterministic convergence rates are derived. It is shown that the estimator enjoys nonasymptotic oracle properties under rank sparsity condition, which is similar to that in Negahban and Wainwright (2011). The technique is applied to a multivariate modification of the famous Conditional Autoregressive Value-at-Risk (CAViaR) model of Engle and Manganelli (2004), which is called Sparse Asymmetric Conditional Value-at-Risk (SAMCVaR). With a dataset consists of stock prices of 230 global financial firms ranging over 2007-2010, we confirm the leverage effect documented in previous studies like Engle and Ng (1993), and furthermore we show that the negative lag return increase the distribution dispersion mostly by lowering the left tail of the distribution, which does not yield the potential for gain. Finally, a nonparametric extension of our method is proposed and applied on Chinese temperature data collected from 159 weather stations for the classification of temperature seasonality patterns.

Chapter 2

Quantile Regression in Risk Calibration

2.1 Introduction

Sufficiently accurate risk measures are needed not only in crisis times. In the last two decades, the world has gone through several financial turmoils, and the financial market is getting riskier and the scale of loss soars. Beside marginal extremes that can shock even a well diversified portfolio, the focus of intensified research in the recent years has been on understanding the interdependence of risk factors and their conditional structure.

The most popular risk measure is the Value-at-Risk (VaR), which is defined as the τ-quantile of the return distribution at time $t+d$ conditioned on the information set $\mathcal{F}_t$:

$$VaR_{t+d}^{\tau} \stackrel{\text{def}}{=} \inf\{x \in \mathbb{R} : \mathrm{P}(X_{t+d} \leq x | \mathcal{F}_t) \geq \tau\}. \tag{2.1.1}$$

Here X_t denotes the asset return and τ is taking values such as 0.05, 0.01 or 0.001 to reflect negative extreme risk.

Extracting information in economic variables to predict VaR brings quantile regression into play here, since VaR is the quantile of the conditional asset return distribution. Engle and Manganelli (2004) propose the nonlinear Conditional Autoregressive Value at Risk (CaViaR) model, which uses (lag) VaR and lag returns. Chernozhukov and Umantsev (2001) propose linear and quadratic time series models for VaR prediction. Kuan et al. (2009) propose the Conditional AutoRegressive Expectile (CARE) model, and argue that expectiles are more sensitive to the scale of losses. These studies and many others apply quantile regression in a prespecified often linear functional form. In a more nonparametric context, Cai and Wang (2008) estimate the conditioned cdf by a double kernel local linear estimator and find the quantile by inverting the cdf. Schaumburg (2011) uses the same technique together

This chapter is published as: Chao, S.-K., Härdle, W. K. and Wang, W. (2014) *Handbook of Financial Econometrics and Statistics*, pp. 1467-1489.

with extreme value theory for VaR prediction. Taylor (2008) proposes Exponentially Weighted Quantile Regression (EWQR) for estimating VaR time series.

The aforementioned studies focus mainly on the VaR estimation for single assets and do not directly take into account the escalated spillover effect in crisis periods. This risk of joint tail events of asset returns has been identified and studied. Further, Brunnermeier and Pedersen (2008) show that the negative feedback effect of a "loss spiral" and a "margin spiral" leads to the joint depreciation of assets prices. It is therefore important to develop risk measures which can quantify the contagion effects of negative extreme event.

Acharya et al. (2010) propose the concept of marginal expected shortfall (MES), which measures the contribution of individual assets to the portfolio expected shortfall. Via an equilibrium argument, the MES is shown to be a predictor to a financial institution's risk contribution. Brownlees and Engle (2010) demonstrate that the MES can be written as a function of volatility, correlation and expectation conditional on tail events. Huang et al. (2011) propose the distress insurance premium (DIP), a measure similar to MES but computed under the risk-neutral probability. This measure can therefore be viewed as the market insurance premium against the event that the portfolio loss exceeds a low level. Adams et al. (2010) construct financial indices on return of insurance companies, commercial banks, investment banks and hedge funds, and use a linear model for the VaRs of the four financial indices to forecast the state-dependent sensitivity VaR (SDSVaR). The risk measures proposed above have some shortcomings though: The computation of DIP is demanding since this involves the simulation of rare events. MES suffers from the scarcity of data because it conditions on a rare event.

In Adrian and Brunnermeier (2011) (henceforth AB), the CoVaR concept of conditional VaR is proposed, which controls the effect of the negative extreme event of some systemically risky financial institutions. Formally, let $C(X_{i,t})$ be some event of a asset i return $X_{i,t}$ at time t and take $X_{j,t}$ as another asset return (e.g. the market index). The $\text{CoVaR}^{\tau}_{j|i,t}$ is defined as the τ-quantile of the conditional probability distribution:

$$\text{P}\left\{ X_{j,t} \leq \text{CoVaR}^{\tau}_{j|i,t} \middle| \, C(X_{i,t}), M_t \right\} = \tau, \tag{2.1.2}$$

where M_t is a vector of market variables defined in Section 2.2.1. The standard CoVaR approach is to set $C(X_{i,t}) = \{X_{i,t} = VaR^{\tau}_{X_{i,t}}\}$. In AB, $X_{j,t}$ is the weekly return which is constructed from a vast data set comprised of all publicly traded commercial banks, broker dealers, insurance companies, and real estate companies in the U.S. Further, AB propose ΔCoVaR (measure of marginal risk contribution) as the difference between $CoVaR^{\tau_1}_{j|i,t}$ and $CoVaR^{\tau_2}_{j|i,t}$, where $\tau_1 = 0.5$ associated with the normal state and $\tau_2 = 0.05$ associated with the financial distress state.

The formulation of this conditional risk measure has several advantages. First, the cloning property: After dividing a systemically risky firm into several clones, the value of CoVaR conditioned on the entire firm does not differ from the one conditioned on one of the clones. Second, the conservativeness. The CoVaR value is more conservative than VaR because it conditions on an extreme event. Third,

CoVaR is endogenously generated and adapted to the varying environment of the market.

The recipe of AB for CoVaR construction is as follows: In a first step one predicts the VaR of an individual asset $X_{i,t}$ through a linear model on market variables:

$$X_{i,t} = \alpha_i + \gamma_i^\top M_{t-1} + \varepsilon_{i,t}, \tag{2.1.3}$$

where $\gamma_i^\top$ means the transpose of γ_i and M_t is a vector of the state variables (see Section 2.2.1). This model is estimated with quantile regression of Koenker and Bassett (1978) to get the coefficients $(\hat{\alpha}_i, \hat{\gamma}_i)$ with $F^{-1}_{\varepsilon_{i,t}}(\tau|M_{t-1}) = 0$. The VaR of asset i is predicted by

$$\widehat{VaR}_{i,t} = \hat{\alpha}_i + \hat{\gamma}_i^\top M_{t-1}. \tag{2.1.4}$$

In a second step one models the asset j return as a linear function of asset return i and market variables M_t:

$$X_{j,t} = \alpha_{j|i} + \beta_{j|i} X_{i,t} + \gamma_{j|i}^\top M_{t-1} + \varepsilon_{j,t}, \tag{2.1.5}$$

again one employs quantile regression and obtains coefficients $(\hat{\alpha}_{j|i}, \hat{\beta}_{j|i}, \hat{\gamma}_{j|i})$. The CoVaR is finally calculated:

$$\widehat{CoVaR}^{AB}_{j|i,t} = \hat{\alpha}_{j|i} + \hat{\beta}_{j|i}\widehat{VaR}_{i,t} + \hat{\gamma}_{j|i}^\top M_{t-1}. \tag{2.1.6}$$

In equation (2.1.5) the variable $X_{i,t}$ influences the return $X_{j,t}$ in a linear fashion. However, the linear parametric model may not be flexible enough to capture the tail dependence between i and j. The linearity of the conditioned quantile curves of X_j on X_i is challenged by the confidence bands of the nonparametric quantile curves, as shown in Figure 2.1.1. The left tail quantile from linear parametric quantile regression (red) lies well outside the confidence band (gray dashed curve) of Härdle and Song (2010). This motivates empirically, that a linear model is not flexible enough for the CoVaR question at hand.

Nonparametric models can be used to account for the nonlinear structure of the conditional quantile, but the challenge for using such models is the curse of dimensionality, as the quantile regression in CoVaR modeling often involves many variables. Thus, we resort to semiparametric partial linear model (PLM) which preserves some flexibility of the nonparametric model while suffers little from the curse of dimensionality.

As an illustration, the VaR/CoVaR of Goldman Sachs (GS) returns are shown, given the returns of Citigroup (C) and S&P500 (SP). S&P500 index return is used as a proxy for the market portfolio return.

Choosing market variables is crucial for the VaR/CoVaR estimation. For the variables representing market states, we follow the most popular choices such as VIX, short term liquidity spread, etc. In particular, the variable we use for real estate companies, is the Dow Jones U.S. real estate index. The data is in daily frequency and spans from August 4, 2006 to August 4, 2011.

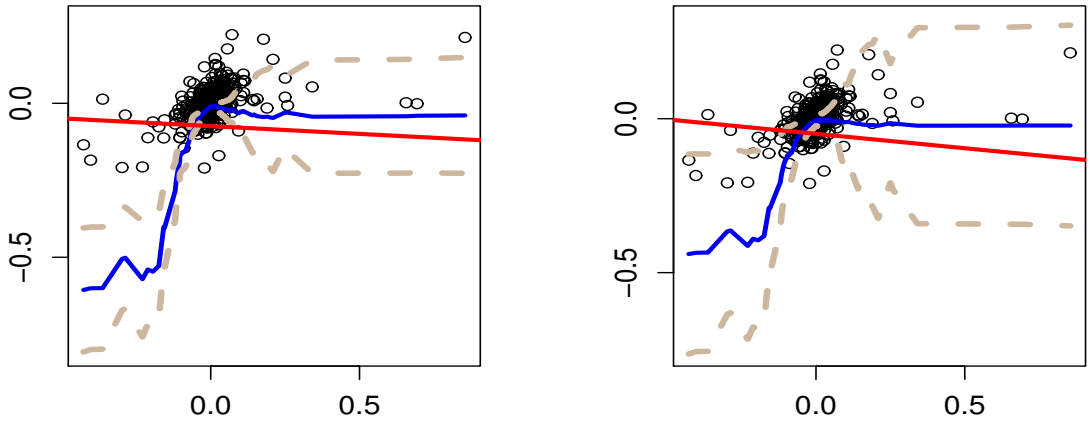


Figure 2.1.1: Goldman Sachs (GS) and Citigroup (C) weekly returns 0.05(left) and 0.1(right) quantile functions. The y-axis is GS daily returns and the x-axis is the C daily returns. The blue curve are the locally linear quantile regression curves (see Appendix A.1). The locally linear quantile regression bandwidth are 0.1026 and 0.0942. The red lines are the linear parametric quantile regression line. The antique white dashed curves are the asymptotic confidence band (see Section A.2) with significance level 0.05. The sample size $N = 546$.

To see if the estimated VaRs/CoVaRs are accurate, we utilize the backtesting procedures described in Berkowitz et al. (2011). We compare three (Co)VaR estimating methods in this study: VaR computed by linear quantile regression on market variables; CoVaR; PLM CoVaR proposed here. The VaR is one-sided interval prediction, the violations (the asset return exceeds estimated VaR/CoVaR) should happen unpredictably if the VaR algorithm is accurate. In other words, the null hypothesis is that the series of violations of VaR is a martingale difference given all the past information. Furthermore, if the time series is autocorrelated, we can reject the null hypothesis of martingale difference right away; therefore, autocorrelation tests can be utilized in this context. The Ljung-Box test is not the most appropriate approach here since it has a too strong null hypothesis (i.i.d. sequence). Thus, we additionally apply the Lobato test. The CaViaR test, which is inspired by the CaViaR model, is proposed and shown to have the best overall performance by Berkowitz et al. (2011) among other alternative tests with an exclusive desk-level data set. To illustrate the VaR/CoVaR performances in the crisis time, we separately apply the CaViaR test to the violations of the whole sample period and to the financial crisis period.

The results show that for the PLM CoVaR of GS given C performs better than the AB and PLM CoVaR given SP during the financial crisis period from mid 2008 to mid 2009. The nonlinearity between GS and C returns may convey information which is incapable to be reflected in the market returns, especially during unstable

market conditions.

In contrast to ΔCoVaR, we use a mathematically more intuitive way to analyze the marginal effect by taking the first order derivative of the quantile function. We call it "marginal contribution of risk" (MCR). Bae et al. (2003) and many others have pointed out the phenomenon of financial contagion across national borders. This motivates us to consider the stock indices of a few developed markets and explore their risk contribution to the global stock market. MCR results show that when the global market condition varies, the source of global market risk can be different. To be more specific, when the global market return is bad, the risk contribution from the U.S. is the largest. On the other hand, during financially stable periods, Hong Kong and Japan are more significant risk contributors than the U.S. to the global market.

This study is organized as follows: Section 2.2 introduces the construction and the estimation of the PLM model of CoVaR. The backtesting methods and our risk contribution measure are also introduced in this section. Section 2.3 presents the Goldman Sachs CoVaR time series and the backtesting procedure results. Section 2.4 presents the conclusion and possible further studies. Appendices describe the detailed estimation and statistical inference procedures used in this study.

2.2 Methodology

Quantile regression is a well-established technique to estimate the conditional quantile function. Koenker and Bassett (1978) focus on the linear functional form. An extension of linear quantile regression is the PLM quantile regression. A partial linear model for the dynamics of assets return quantile is constructed in this section. The construction is justified by a linearity test based on a conservative uniform confidence band proposed in Härdle and Song (2010). For more details on semiparametric modeling and PLM, we refer to Härdle et al. (2004) and Härdle et al. (2000).

The backtesting procedure is done via the CaViaR test. Finally, the methodology of MCR is introduced, which is an intuitive marginal risk contribution measure. We will apply the method to a data set of global market indices in developed countries.

2.2.1 Constructing Partial Linear Model (PLM) for CoVaR

Recall how the CoVaR is constructed:

$$\widehat{VaR}_{i,t} = \hat{\alpha}_i + \hat{\gamma}_i M_{t-1},$$
$$\widehat{CoVaR}^{AB}_{j|i,t} = \hat{\alpha}_{j|i} + \hat{\beta}_{j|i}\widehat{VaR}_{i,t} + \hat{\gamma}^{\top}_{j|i} M_{t-1}.$$

where $(\hat{\alpha}_i, \hat{\gamma}_i)$ and $(\hat{\alpha}_{j|i}, \hat{\beta}_{j|i}, \hat{\gamma}_{j|i})$ are estimated from a linear model using standard linear quantile regression.

We have motivated the need for more general functional forms for the quantile curve. We therefore relax the model to a non- or semiparametric model. The market

variable M_t is multidimensional and the data frequency here is daily. The following key variables are entering our analysis:

1. VIX: Measuring the model-free implied volatility of the market. This index is known as the "fear gauge" of investors. The historical data can be found on the Chicago Board Options Exchange's website.

2. Short term liquidity spread: Measuring short-term liquidity risk by the difference between the three-month treasury repo rate and the three-month treasury bill rate. The repo data is from the Bloomberg database and the treasury bill rate data is from the Federal Reserve Board H.15.

3. The daily change in the three-month treasury bill rate: AB find that the changes have better explanatory power than the levels for the negative tail behavior of asset returns.

4. The change in the slope of the yield curve: The slope is defined by the difference of the ten-year treasury rate from the three-month treasury bill rate.

5. The change in the credit spread between 10 years BAA-rated bonds and the 10 years treasury rate.

6. The daily Dow Jones U.S. Real Estate index returns: The index reflects the information of lease rates, vacancies, property development and transactions of real estates in the U.S.

7. The daily S&P500 index returns: The approximate of the theoretical market portfolio returns.

The variables 3, 4, 5 are from the Federal Reserve Board H.15 and the data of 6 and 7 are from Yahoo Finance.

First we conduct a statistical check of the linearity between GS return and the market variables using the confidence band as constructed in Appendix A.2. As shown in Figure 2.2.1 (a) and 2.2.2 (b), except for some ignorable outsiders, the linear quantile regression line lies in the LLQR asymptotic confidence band.

On the other hand, there is nonlinearity between two individual assets X_i and X_j. To illustrate this, we regress X_j on M_t, and then take the residuals and regress them on X_i. Again the $X_{j,t}$ is GS daily return and X_i is C daily return. The result is shown in Figure 2.2.3. The linear QR line (red) lies well outside the LLQR confidence band (magenta) when the C return is negative. The linear quantile regression line is fairly flat. The risk of using a linear model is obvious in this figure: the linear regression can "average out" the humped relation of the underlying structure (blue), and therefore imply a model risk in estimation.

Based on the results of the linearity tests above, we construct a PLM model:

$$X_{i,t} = \alpha_i + \gamma_i^\top M_{t-1} + \varepsilon_{i,t}, \tag{2.2.1}$$

$$X_{j,t} = \tilde{\alpha}_{j|i} + \tilde{\beta}_{j|i}^\top M_{t-1} + l_{j|i}(X_{i,t}) + \varepsilon_{j,t}, \tag{2.2.2}$$

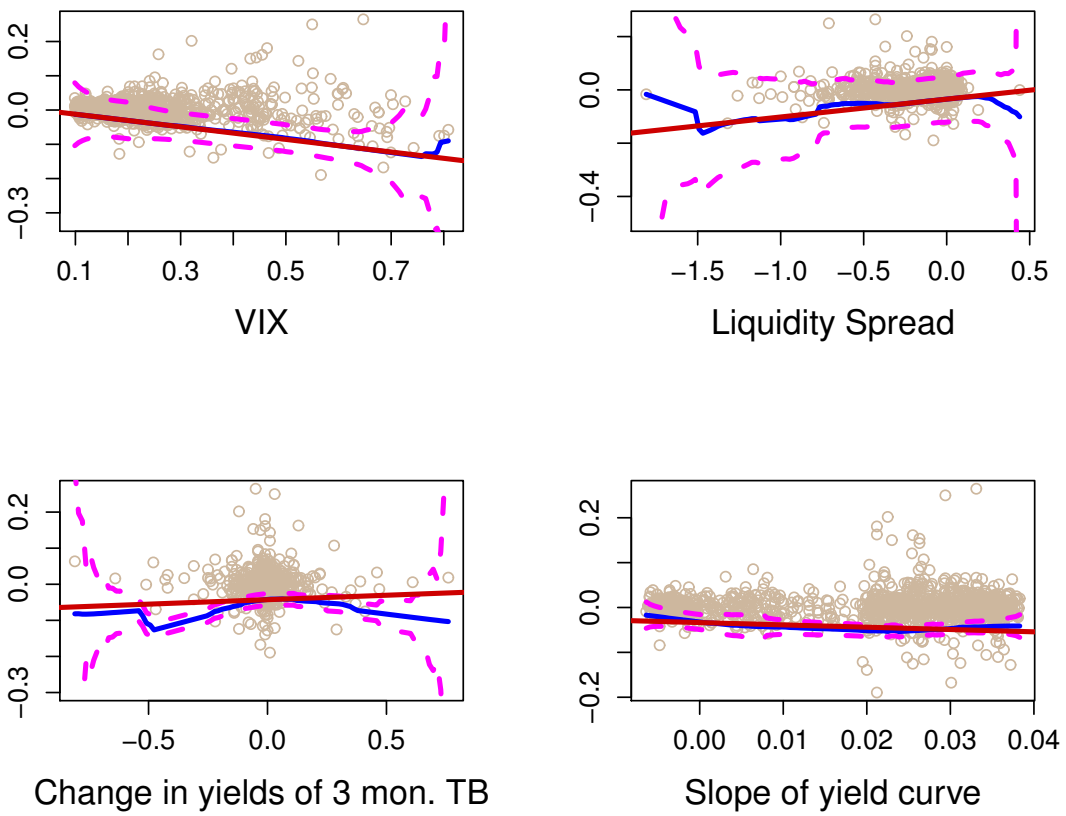
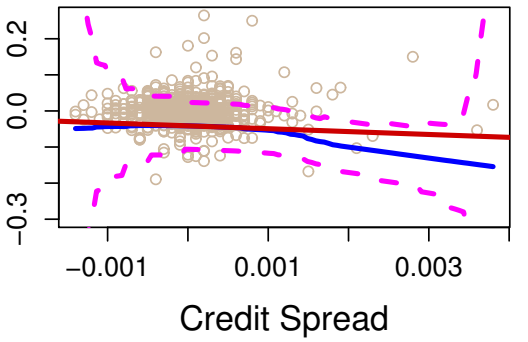
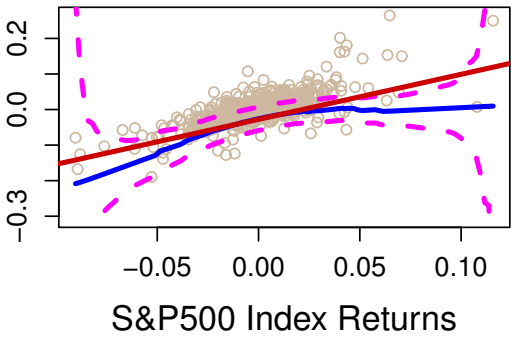


Figure 2.2.1: The scatter plots of GS daily returns to the 7 market variables with the LLQR curves. The bandwidths are selected by the method described in Appendix A.1. The LLQR bandwidths are 0.1101, 0.1668, 0.2449, 0.0053, 0.0088, 0.0295 and 0.0569. The data period is from August 4, 2006 to August 4, 2011. $N = 1260$. $\tau = 0.05$





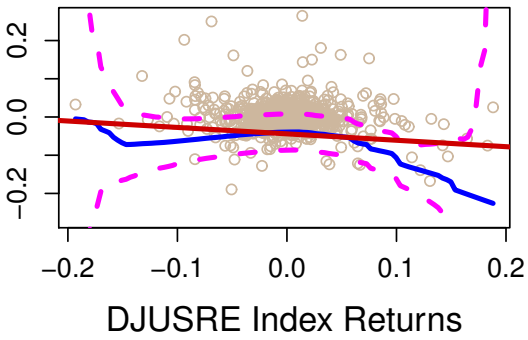


Figure 2.2.2: (Continued from Figure 2.2.1)

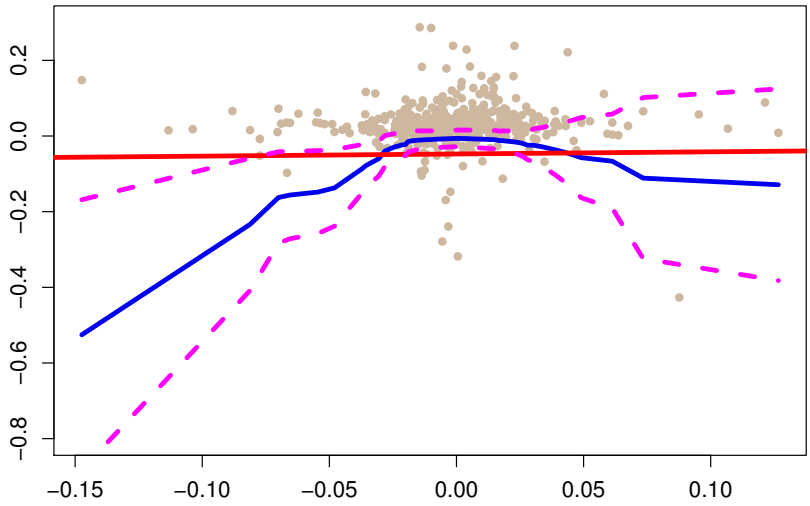


Figure 2.2.3: The nonparametric part $\hat{l}_{GS|C}(\cdot)$ of the PLM estimation. The y-axis is the GS daily returns. The x-axis is the C daily returns. The blue curve is the LLQR quantile curve. The red line is the linear parametric quantile line. The magenta dashed curves are the asymptotic confidence band with significance level 0.05. The data is from June 25, 2008 to December 23, 2009. 378 observations. Bandwidth =0.1255. $\tau = 0.05$.

where $X_{i,t}, X_{j,t}$ are asset returns of i, j firms. M_t is a vector of market variables at time t as introduced before. If i=S&P500, M_t is set to consist of the first 6 market variables only. Notice the variable $X_{i,t}$ enter the equation (2.2.2) nonlinearly.

Applying the algorithm of Koenker and Bassett (1978) to (2.2.1) and the process described in Appendix A.3 to equation (2.2.2), we get $\{\hat{\alpha}_i, \hat{\gamma}_i\}$ and $\{\hat{\alpha}_{j|i}, \hat{\beta}_i, \hat{l}(\cdot)\}$ with $F^{-1}_{\varepsilon_{i,t}}(\tau|M_{t-1}) = 0$ for (2.2.1) and $F^{-1}_{\varepsilon_{j,t}}(\tau|M_{t-1}, X_{i,t}) = 0$ for (2.2.2). Finally, we estimate the PLM $CoVaR_{j|i,t}$ by

$$\widehat{VaR}_{i,t} = \hat{\alpha}_i + \hat{\gamma}_i^\top M_{t-1}, \tag{2.2.3}$$

$$\widehat{CoVaR}^{PLM}_{j|i,t} = \hat{\tilde{\alpha}}_{j|i} + \hat{\tilde{\beta}}_j^\top M_{t-1} + \hat{l}_{j|i}(\widehat{VaR}_{i,t}). \tag{2.2.4}$$

2.2.2 Backtesting

The goal of the backtesting procedure is to check if the VaR/CoVaR is accurate enough so that managerial decisions can be made based on them. The VaR forecast is a (one-sided) interval forecast. If the VaR algorithm is correct, then the violations should be unpredictable, after using all the past information. Formally, if we define

the violation time series as

$$I_t = \begin{cases} 1, & \text{if } X_t < \widehat{VaR_t^\tau}; \\ 0, & \text{otherwise.} \end{cases}$$

Where $\widehat{VaR_t^\tau}$ can be replaced by $\widehat{CoVaR_t^\tau}$ in the case of CoVaR. I_t should form a sequence of martingale difference.

There is a large literature on martingale difference tests. We adopt Ljung-Box test, Lobato test and the CaViaR test. The Ljung-Box test and Lobato test aim to check whether the time series is autocorrelated. If the time series is autocorrelated, then we reject of course the hypothesis that the time series is a martingale difference.

Particularly, let $\hat{\rho}_k$ be the estimated autocorrelation of lag k of the sequence of violation $\{I_t\}$ and n be the length of the time series. The Ljung-Box test statistics is:

$$\mathrm{LB}(m) = n(n+2)\sum_{k=1}^{m} \frac{\hat{\rho}_k^2}{n-k} \xrightarrow{\mathcal{L}} \chi(m), \tag{2.2.5}$$

as $n \to \infty$.

This test is too strong though in the sense that the asymptotic distribution is derived based on the i.i.d. assumption. A modified Box-Pierce test is proposed by Lobato et al. (2001), who also consider the test of no autocorrelation, but their test is more robust to the correlation of higher (greater than the first) moments. (Autocorrelation in higher moments does not contradict with the martingale difference hypothesis.) The test statistics is given by

$$\mathrm{L}(m) = n\sum_{k=1}^{m} \frac{\hat{\rho}_k^2}{\hat{v}_{kk}} \xrightarrow{\mathcal{L}} \chi(m),$$

as $n \to \infty$, where

$$\hat{v}_{kk} = \frac{\frac{1}{n}\sum_{i=1}^{n-k}(y_i - \bar{y})^2(y_{i+k} - \bar{y})^2}{\{\frac{1}{N}\sum_{i=1}^{n}(y_i - \bar{y})^2\}^2}.$$

The CaViaR test, proposed by Berkowitz et al. (2011), is based on the idea that if the sequence of violation is a martingale difference, there ought to be no correlation between any function of the past variables and the current violation. One way to test this uncorrelatedness is through a linear model. The model is:

$$I_t = \alpha + \beta_1 I_{t-1} + \beta_2 VaR_t + u_t,$$

where VaR_t can be replaced by $CoVaR_t$ in the case of conditional VaR. The residual u_t follows a Logistic distribution since I_t is binary. We get the estimates of the coefficients $(\hat{\beta}_1, \hat{\beta}_2)^\top$. Therefore the null hypothesis is $\hat{\beta}_1 = \hat{\beta}_2 = 0$. This hypothesis can be tested by Wald's test.

We set $m = 1$ or 5 for the Ljung-Box and Lobato tests. For the CaViaR test, two data periods are considered separately. The first is the overall data from August 4, 2006 to August 4, 2011. The second is the data from August 4, 2008 to August 4, 2009, the period when the financial market reached its bottom. By separately testing the two periods, we can gain more insights into the PLM model.

2.2.3 Risk contribution measure

The risk contribution of one firm to the market is one of the top concerns among central bankers. The regulator can restrict the risky behaviors of the financial institution with high risk contribution to the market, and reduce the institution's incentive to take more risk. AB propose the idea of ΔCoVaR, which is defined by

$$\Delta CoVaR^{\tau}_{j|i,t} = CoVaR^{\tau}_{j|i,t} - CoVaR^{0.5}_{j|i,t}. \tag{2.2.6}$$

where $\text{CoVaR}^{\tau}_{j|i,t}$ is defined as in the introduction. j, i represent the financial system and an individual asset. $\tau = 0.5$ corresponds to the normal state of the individual asset i. This is essentially a sensitivity measure quantifying the effect to the financial system from the occurrence of a tail event of asset X_i.

In this study we adopt a mathematically intuitive way to measure the marginal effect by searching the first order derivative of the quantile function. Because the spillover effect from stock market to stock market has already got much attention, it is important to investigate the risk contribution of a local market to the global stock market. The estimation is conducted as follows:

First, one estimates the following model nonparametrically:

$$X_{j,t} = f_j^{0.05}(X_t) + \varepsilon_j, \tag{2.2.7}$$

The quantile function $f_j^{0.05}(\cdot)$ is estimated with local linear quantile regression with $\tau = 0.05$, described with more details in Appendix A.1. X_j is the weekly return of the stock index of an individual country and X is the weekly return of the global stock market.

Second, with $\hat{f}_j^{0.05}(\cdot)$, we compute the "marginal contribution of risk"(MCR) of institution j by

$$MCR^{\tau}_j = \frac{\partial \hat{f}_j^{0.05}(x)}{\partial x}\Big|_{x=\hat{F}_X^{-1}(\tau_k)}, \tag{2.2.8}$$

where $\hat{F}^{-1}(\tau_k)$ is a consistent estimator of the τ_k quantile of the global market return, and it can be estimated by regressing X_t on the time trend. We put $k = 1, 2$ with $\tau_1 = 0.5$ and $\tau_2 = 0.05$. The quantity (2.2.8) is similar to the MES proposed by Acharya et al. (2010) in the sense that the conditioned event belongs to the information set of the market return, but we reformulate it in the VaR framework instead of the expected shortfall framework.

There are some properties of the MCR to be described further. First, τ_k determines the condition of the global stock market. This allows us to explore the risk contribution from the index j to the global market given different global market status. Second, the higher the value of MCR, the more risk factor j imposes on the market in terms of risk. Third, since the function $f_j^{0.05}(\cdot)$ is estimated by LLQR, the quantile curve is locally linear, and therefore the local first order derivative is straightforward to compute.

We choose indices j=S&P500, NIKKEI225, FTSE100, DAX30, CAC40, Heng Seng as the approximate of the market returns of each developed country or market.

The global market is approximated by the MSCI World (developed countries) market index. The data is weekly from April 11, 2004 to April 11, 2011 and $\tau = 0.05$

2.3 Results

2.3.1 CoVaR estimation

The estimation results of VaR/CoVaR are shown in this section. We compute three types of VaR/CoVaR of GS, with a moving window size of 126 business days and $\tau = 0.05$.

First, the VaR of GS is estimated:

$$\widehat{VaR}_{GS,t} = \hat{\alpha}_{GS} + \hat{\gamma}_{GS}^{\top} M_{t-1}, \tag{2.3.1}$$

using linear quantile regression, and $M_t \in \mathbb{R}^7$ is introduced in Section 2.2.1.

Second, the CoVaR of GS given C returns is estimated:

$$\widehat{VaR}_{C,t} = \hat{\alpha}_{C} + \hat{\gamma}_{C}^{\top} M_{t-1}; \tag{2.3.2}$$

$$\widehat{CoVaR}^{AB}_{GS|C,t} = \hat{\alpha}_{GS|C} + \hat{\beta}_{GS|C}\widehat{VaR}_{C,t} + \hat{\gamma}_{GS|C}^{\top} M_{t-1}. \tag{2.3.3}$$

If the SP replaces C, the estimates are generated from

$$\widehat{VaR}_{SP,t} = \hat{\alpha}_{SP} + \hat{\gamma}_{SP}^{\top} \widetilde{M}_{t-1}; \tag{2.3.4}$$

$$\widehat{CoVaR}^{AB}_{GS|SP,t} = \hat{\alpha}_{GS|SP} + \hat{\beta}_{GS|SP}\widehat{VaR}_{SP,t} + \hat{\gamma}_{GS|SP}^{\top} \widetilde{M}_{t-1}, \tag{2.3.5}$$

where $\widetilde{M}_t \in \mathbb{R}^6$ is the vector of market variables without the market portfolio return.

Third, the PLM CoVaR is generated:

$$\widehat{VaR}_{C,t} = \hat{\alpha}_{C} + \hat{\gamma}_{C}^{\top} M_{t-1}; \tag{2.3.6}$$

$$\widehat{CoVaR}^{PLM}_{GS|C,t} = \hat{\tilde{\alpha}}_{GS|C} + \hat{\tilde{\beta}}_{GS|C}^{\top} M_{t-1} + \hat{l}_{GS|C}(\widehat{VaR}_{C,t}). \tag{2.3.7}$$

If SP replaces C:

$$\widehat{VaR}_{SP,t} = \hat{\alpha}_{SP} + \hat{\gamma}_{SP}^{\top} \widetilde{M}_{t-1}; \tag{2.3.8}$$

$$\widehat{CoVaR}^{PLM}_{GS|SP,t} = \hat{\tilde{\alpha}}_{GS|SP} + \hat{\tilde{\beta}}_{GS|SP}^{\top} \widetilde{M}_{t-1} + \hat{l}_{GS|SP}(\widehat{VaR}_{SP,t}). \tag{2.3.9}$$

The coefficients in (2.3.1), (2.3.2), (2.3.3), (2.3.4), (2.3.5), (2.3.6) and (2.3.8) are estimated from the linear quantile regression and those in (2.3.7) and (2.3.9) are estimated from the method described in Appendix A.3.

Figure 2.3.1 shows the $\widehat{VaR}_{GS,t}$ sequence. The VaR forecasts (red) seem to form a lower cover of the GS returns (blue). This suggests that the market variables M_t have some predictive power for the left tail quantile of the GS return distribution. Figure 2.3.2 shows the sequences $\widehat{CoVaR}^{AB}_{GS|SP,t}$ (cyan) and $\widehat{CoVaR}^{PLM}_{GS|C,t}$ (light green). As

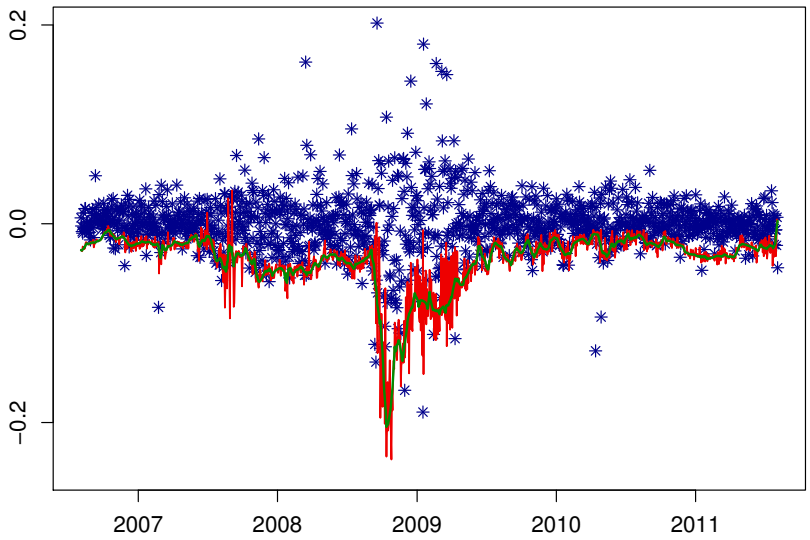


Figure 2.3.1: The $\widehat{VaR}_{GS,t}$. The red line is the $\widehat{VaR}_{GS,t}$ and blue stars are daily returns of GS. The dark green curve is the meadian smoother of the $\widehat{VaR}_{GS,t}$ curve with h=2.75. $\tau = 0.05$. The window size is 252 days.

the time series of the estimates is too volatile, we smooth it further by the median LLQR. The two estimates are similar as the market state is stable, but during the period of financial instability (from mid 2008 to mid 2009), the two estimates have different behavior. The performance of these estimates are evaluated by backtesting procedure in Section 2.3.2.

Table 2.3.1 shows the summary statistics of the VaR/CoVaR estimates. The first three rows show the summary statistics of $\widehat{VaR}_{GS,t}$, $\widehat{VaR}_{C,t}$ and $\widehat{VaR}_{SP,t}$. The $\widehat{VaR}_{GS,t}$ has lower mean and higher standard deviation than the other two. Particularly during 2008 to 2009, the standard deviation of the GS VaR is twice as much as the other two. The mean and standard deviation of the $\widehat{VaR}_{C,t}$ and $\widehat{VaR}_{SP,t}$ are rather similar. The last four rows show the summary statistics of $\widehat{CoVaR}^{PLM}_{GS|C,t}$, $\widehat{CoVaR}^{AB}_{GS|C,t}$, $\widehat{CoVaR}^{PLM}_{GS|SP,t}$ and $\widehat{CoVaR}^{AB}_{GS|SP,t}$. This shows that the CoVaR obtaining from the AB model has smaller mean but greater standard deviation than the CoVaR obtaining from PLM model.

Figure 2.3.3 shows the bandwidth sequence of the nonparametric part of the PLM estimation. The bandwidth varies with time. Before mid 2007, the bandwidth sequence is stably jumping around 0.2. After that the sequence becomes very volatile. This may have something to do with the rising systemic risk.

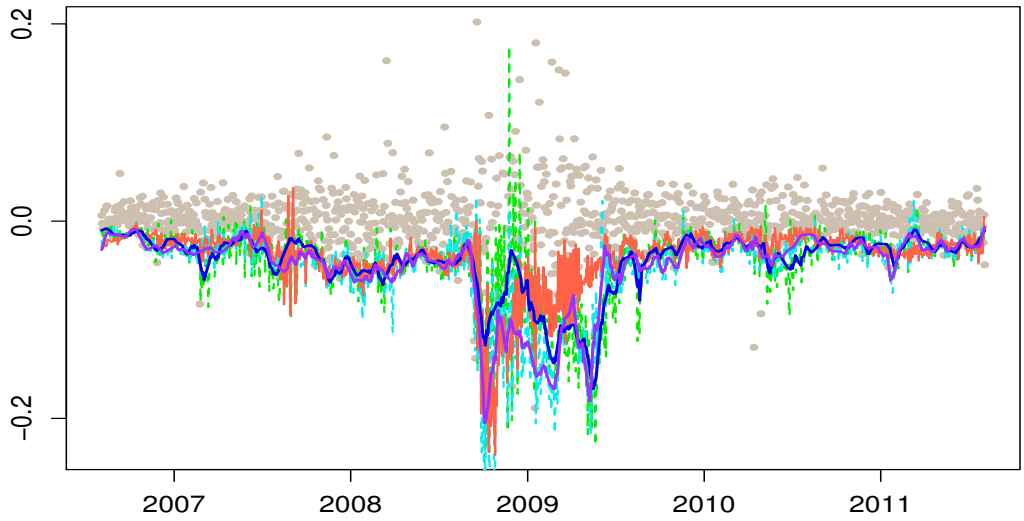


Figure 2.3.2: The CoVaR of GS given the VaR of C. The gray dots are daily returns of GS. The light green dashed curve is the $\widehat{CoVaR}^{PLM}_{GS|C,t}$. The blue curve is the median LLQR smoother of the light green dashed curve with $h = 3.19$. The cyan dashed curve is the $\widehat{CoVaR}^{AB}_{GS|C,t}$. The purple curve is the median LLQR smoother of the cyan dashed curve with $h = 3.90$. The red curve is the $\widehat{VaR}_{GS,t}$. $\tau = 0.05$. The moving window size is 126 days.

	mean-overall	sd-overall	mean-crisis	sd-crisis
$\widehat{VaR}_{GS,t}$	-3.66	3.08	-7.43	4.76
$\widehat{VaR}_{C,t}$	-2.63	1.67	-4.62	2.25
$\widehat{VaR}_{SP,t}$	-2.09	1.57	-3.88	2.24
$\widehat{CoVaR}^{PLM}_{GS\|C,t}$	-4.26	3.84	-8.79	5.97
$\widehat{CoVaR}^{AB}_{GS\|C,t}$	-4.60	4.30	-10.36	6.32
$\widehat{CoVaR}^{PLM}_{GS\|SP,t}$	-3.86	3.30	-8.20	4.69
$\widehat{CoVaR}^{AB}_{GS\|SP,t}$	-5.81	4.56	-12.65	5.56

Table 2.3.1: VaR/CoVaR summary statistics. The overall period is from August 4, 2006 to August 4, 2011. The crisis period is from August 4, 2008 to August 4, 2009. The numbers in the table are scaled up by 10^2.

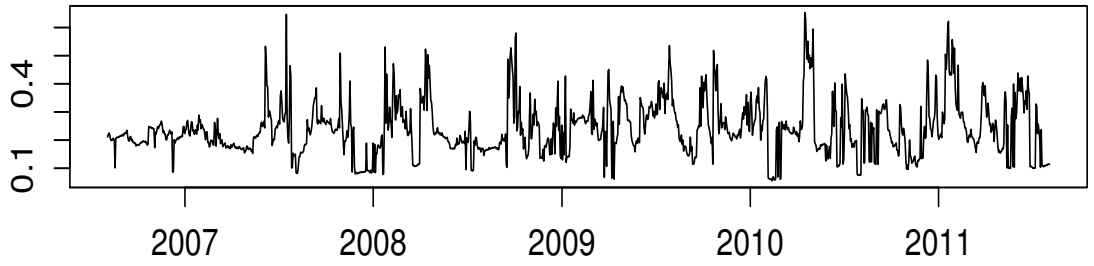


Figure 2.3.3: LLQR bandwidth in the moving daily estimation of $\widehat{CoVaR}^{PLM}_{GS|C,t}$. The average bandwidth is 0.24.

2.3.2 Backtesting

For the evaluation of the CoVaR models, we resort to the backtesting procedure described in Section 2.2.2. In order to perform the backtesting procedure, the sequences $\{I_t\}$ (defined in Section 2.2.2) have to be computed for all VaR/CoVaR estimates. Figure 2.3.4 shows the timings of the violations $\{t : I_t = 1\}$ of $\widehat{CoVaR}^{PLM}_{GS|C,t}$, $\widehat{CoVaR}^{AB}_{GS|C,t}$ and $\widehat{VaR}_{GS,t}$. This figure shows the total number of violations of PLM CoVaR and CoVaR are similar, while $\widehat{VaR}_{GS,t}$ has more violations than the both. The $\widehat{VaR}_{GS,t}$ has a few clusters of violations in both financial stable and unstable periods. This may result from the failure $\widehat{VaR}_{GS,t}$ to adapt for the negative shocks. The violations of $\widehat{CoVaR}^{PLM}_{GS|C,t}$ are more evenly distributed. The violations of $\widehat{CoVaR}^{AB}_{GS|C,t}$ have large clusters during financially stable period, while the violation during financial crisis period is meager. This contrast suggests that $\widehat{CoVaR}^{AB}_{GS|C,t}$ tend to overreact, as it is slack during the stable period but is too tight during the unstable period.

Figure 2.3.5 shows the timings of the violations $\{t : I_t = 1\}$ of $\widehat{CoVaR}^{PLM}_{GS|SP,t}$, $\widehat{CoVaR}^{AB}_{GS|SP,t}$ and $\widehat{VaR}_{GS,t}$. The overall number of violations of $\widehat{CoVaR}^{PLM}_{GS|SP,t}$ is more than that of $\widehat{VaR}_{GS,t}$, and it has many clusters. $\widehat{CoVaR}^{PLM}_{GS|SP,t}$ behaves differently from $\widehat{CoVaR}^{PLM}_{GS|C,t}$. The SP may not be more informative than C, though the efficient market hypothesis suggests so. The violation of $\widehat{CoVaR}^{AB}_{GS|SP,t}$ is fewer than the other two measures, and the clustering is not significant.

The backtesting procedure is performed separately for each sequence of $\{I_t\}$. The null hypothesis is that each sequence $\{I_t\}$ forms a series of martingale difference. Six different tests are applied for each $\{I_t\}$: Ljung-Box tests with lags 1 and 5, Lobato test with lags 1 and 5 and finally the CaViaR test with two data periods: overall and crisis period.

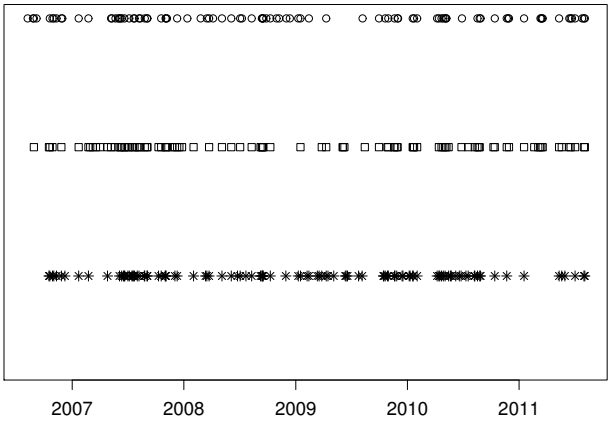


Figure 2.3.4: The timings of violations $\{t : I_t = 1\}$. The top circles are the violations of the $\widehat{CoVaR}^{PLM}_{GS|C,t}$, totally 95 violations. The middle squares are the violations of $\widehat{CoVaR}^{AB}_{GS|C,t}$, totally 98 violations. The bottom stars are the violations of $\widehat{VaR}_{GS,t}$, totally 109 violations. Overall data $N = 1260$.

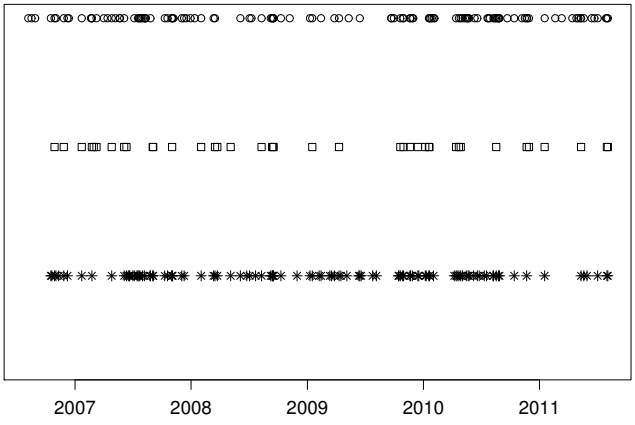


Figure 2.3.5: The timings of violations $\{t : I_t = 1\}$. The top circles are the violations of $\widehat{CoVaR}^{PLM}_{GS|SP,t}$, totally 123 violations. The middle squares are the violations of $\widehat{CoVaR}^{AB}_{GS|SP,t}$, totally 39 violations. The bottom stars are the violations of $\widehat{VaR}_{GS,t}$, totally 109 violations. Overall data $N = 1260$.

The result is shown in Table 2.3.2. First, in Panel 1 of Table 2.3.2, the $\widehat{VaR}_{GS,t}$ is rejected by the LB(5) test and the two CaViaR tests. This shows that a linear

quantile regression on the seven market variables may not give accurate estimates, in the sense that the violation $\{I_t\}$ of $\widehat{VaR}_{GS,t}$ does not form a martingale sequence. Next we turn to the $\widehat{CoVaR}^{AB}_{GS|SP,t}$ and $\widehat{CoVaR}^{PLM}_{GS|SP,t}$. In Panel 2, the low p-values of the two CaViaR tests show that both the AB model and PLM model conditioned on SP are rejected, though the p-value of the AB model almost reaches the 5% significant level. In particular, the $\widehat{CoVaR}^{PLM}_{GS|SP,t}$ is rejected by the L(5) and LB(5) tests. Both the parametric and semiparametric models fail with this choice of variable. This suggests that the market return does not provide enough information in risk measurement.

We therefore need more informative variables. Panel 3 of Table 2.3.2 illustrates this by using C daily returns, which may contain information not revealed in the market and improve the performance of the estimates. The $\widehat{CoVaR}^{AB}_{GS|C,t}$ is rejected by the two CaViaR tests and the LB(1) test with 0.1% and 5% significant level. However, $\widehat{CoVaR}^{PLM}_{GS|C,t}$ is not rejected by the CaViaR-crisis test. This implies that the nonparametric part in the PLM model captures the nonlinear effect of C returns to GS returns, which can lead to better risk-measuring performance.

Measure	LB(1)	LB(5)	L(1)	L(5)	CaViaR-overall	CaViaR-crisis
			Panel 1			
$\widehat{VaR}_{GS,t}$	0.3449	0.0253*	0.3931	0.1310	1.265×10^{-6}***	0.0024**
			Panel 2			
$\widehat{CoVaR}^{AB}_{GS\|SP,t}$	0.0869	0.2059	0.2684	0.6586	8.716×10^{-7}***	0.0424*
$\widehat{CoVaR}^{PLM}_{GS\|SP,t}$	0.0518	0.0006***	0.0999	0.0117*	2.2×10^{-16}***	0.0019**
			Panel 3			
$\widehat{CoVaR}^{AB}_{GS\|C,t}$	0.0489*	0.2143	0.1201	0.4335	3.378×10^{-9}***	0.0001***
$\widehat{CoVaR}^{PLM}_{GS\|C,t}$	0.8109	0.0251*	0.8162	0.2306	2.946×10^{-9}***	0.0535

*, ** and *** denote significance at the 5, 1 and 0.1 percent levels.

Table 2.3.2: Goldman Sachs VaR/CoVaR backtesting p-values. The overall period is from August 4, 2006 to August 4, 2011. The crisis period is from August 4, 2008 to August 4, 2009. LB(1) and LB(5) are the Ljung-Box tests of lags 1 and 5. L(1) and L(5) are the Lobato tests of lags 1 and 5. CaViaR-overall and CaViaR-crisis are two CaViaR tests described in Section 2.2.2 applied on the two data periods.

2.3.3 Global risk contribution

In this section we present the MCR (defined in Section 2.2.3), which measures the marginal risk contribution of risk factors. We choose $\tau_1 = 0.5$, associated to the normal (median) state and $\tau_2 = 0.05$, associated to an negative extreme state. Figure 2.3.6 shows the $MCR_j^{\tau_1}$ from local markets j to the global market. When the MSCI World is (hypothetically) at its normal state, one concludes that the Heng Seng in normal times contributes the most to the MSCI World at all times. The

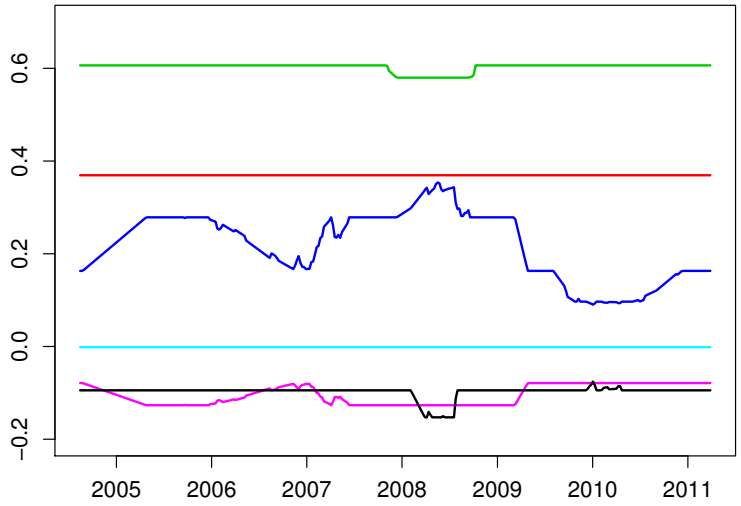


Figure 2.3.6: The $MCR_j^{\tau_1}$, $\tau = 0.5$. j:CAC, FTSE, DAX, Heng Seng, S&P500 and NIKKEI225. The global market return is approximated by MSCI World.

NIKKEI225 places second; the contribution from S&P500 varies most with the time; the risk contribution from DAX30 is nearly zero. The contribution from CAC40 and FTSE100 are negative.

Assuming that the MSCI World is at its bad state ($\tau_2 = 0.05$), the $MCR_j^{\tau_2}$ differs from $MCR_j^{\tau_1}$, see Figure 2.3.7. One sees that the S&P500 imposes more pressure on the world economy than the other countries. Especially during the financial crisis of 2008 and 2009. The contribution from Heng Seng is no longer of the same significance. The three European markets are relatively stable.

This analysis suggests that the risk contribution from individual stock market varies a lot with the state of global economy.

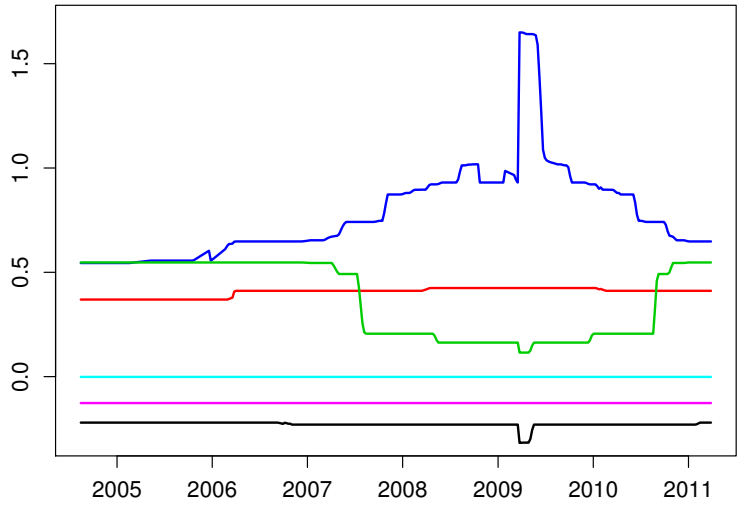


Figure 2.3.7: The $MCR_j^{\tau_2}$, $\tau = 0.05$. j:CAC, FTSE, DAX, Heng Seng, S&P500 and NIKKEI225. The global market return is approximated by MSCI World.

2.4 Conclusion

In this study we construct a PLM model for the CoVaR, and we compare it to the AB model by backtesting. Results show that PLM CoVaR is preferable especially during a crisis period. The study of the MCR reveals the fact that the risk from each country can vary with the state of global economy.

As an illustration, we only study the Goldman Sachs conditional VaR with Citigroup and S&P500 as conditioned risk sources. In practice, we need to choose variables. In Hautsch et al. (2014), the Least Absolute Shrinkage and Selection Operator (LASSO) techniques is used to determine the most relevant systemic risk sources from a pool of financial institutions. A VAR (Vector Autoregression) model may be also suitable for capturing the asset dynamics, but the estimation may be more involved. We may include other firm specific variables such as corporate bond yields as these variables can bear other information which is not included in the stock returns or stock indices.

Chapter 3

Confidence Corridors for Generalized Quantile Regression

3.1 Introduction

Mean regression analysis is a widely used tool in statistical inference for curves. It focuses on the center of the conditional distribution, given d-dimensional covariates with $d \geq 1$. In a variety of applications though the interest is more in tail events, or even tail event curves such as the conditional quantile function. Applications with a specific demand in tail event curve analysis include finance, climate analysis, labor economics and systemic risk management.

Tail event curves have one thing in common: they describe the likeliness of extreme events conditional on the covariate $\boldsymbol{X}$. A traditional way of defining such a tail event curve is by translating "likeliness" with "probability" leading to conditional quantile curves. Extreme events may alternatively be defined through conditional moment behaviour leading to more general tail descriptions as studied by Newey and Powell (1987) and Jones (1994). We employ this more general definition of generalized quantile regression (GQR), which includes, for instance, expectile curves and study statistical inference of GQR curves through confidence corridors.

In applications parametric forms are frequently used because of practical numerical reasons. Efficient algorithms are available for estimating the corresponding curves. However, the "monocular view" of parametric inference has turned out to be too restrictive. This observation prompts the necessity of checking the functional form of GQR curves. Such a check may be based on testing different kinds of variation between a hypothesized (parametric) model and a smooth alternative GQR. Such an approach though involves either an explicit estimate of the bias or a presmoothing of the "null model". In this paper we pursue the Kolmogorov-Smirnov type of approach, that is, employing the maximal deviation between the null and the smooth GQR curve as a test statistic. Such a model check has the advantage

This chapter is published as working paper: Chao, S.-K., Proksch, K., Dette, H. and Härdle, W. K. (2014). Confidence Corridors for Multivariate Generalized Quantile Regression. *SFB 649 Discussion Paper*, 2014-028, Humboldt-Universität zu Berlin.

that it may be displayed graphically as a confidence corridor (CC; also called "simultaneous confidence band" or "uniform confidence band/region") but has been considered so far only for univariate covariates. The basic technique for constructing CC of this type is extreme value theory for the sup-norm of an appropriately centered nonparametric estimate of the quantile curve.

Confidence corridors with one-dimensional predictor were developed under various settings. Classical one-dimensional results are confidence bands constructed for histogram estimators by Smirnov (1950) or more general one-dimensional kernel density estimators by Bickel and Rosenblatt (1973). The results were extended to a univariate nonparametric mean regression setting by Johnston (1982), followed by Härdle (1989) who derived CCs for one-dimensional kernel M-estimators. Claeskens and Van Keilegom (2003) proposed uniform confidence bands and a bootstrap procedure for regression curves and their derivatives.

In recent years, the growth of the literature body shows no sign of decelerating. In the same spirit of Härdle (1989), Härdle and Song (2010) and Guo and Härdle (2012) constructed uniform confidence bands for local constant quantile and expectile curves. Fan and Liu (2013) proposed an integrated approach for building simultaneous confidence band that covers semiparametric models. Giné and Nickl (2010) investigated adaptive density estimation based on linear wavelet and kernel density estimators and Lounici and Nickl (2011) extended the framework of Bissantz et al. (2007) to adaptive deconvolution density estimation. Bootstrap procedures are proposed as a remedy for the poor coverage performance of asymptotic confidence corridors. For example, the bootstrap for the density estimator is proposed in Hall (1991) and Mojirsheibani (2012), and for local constant quantile estimators in Song et al. (2012).

However, only recently progress has been achieved in the construction of confidence bands for regression estimates with a multivariate predictor. Hall and Horowitz (2013) derived an expansion for the bootstrap bias and established a somewhat different way to construct confidence bands without the use of extreme value theory. Their bands are uniform with respect to a fixed but unspecified portion (smaller than one) of points in a possibly multidimensional set in contrast to the classical approach where uniformity is achieved on the complete set considered. Proksch et al. (2015) proposed multivariate confidence bands for convolution type inverse regression models with fixed design.

To the best of our knowledge, the classical Smirnov-Bickel-Rosenblatt type confidence corridors are not available for multivariate GQR or mean regression with random design.

In this work we go beyond the earlier studies in three aspects. First, we extend the applicability of the CC to d-dimensional covariates with $d > 1$. Second, we present a more general approach covering not only quantile or mean curves but also GQR curves that are defined via a minimum contrast principle. Third, we propose a bootstrap procedure and we show numerically its improvement in the coverage accuracy as compared to the asymptotic approach.

Our asymptotic results, which describe the maximal absolute deviation of gen-

eralized quantile estimators, can not only be used to derive a goodness-of-fit test in quantile and expectile regression, but they are also applicable in testing the quantile treatment effect and stochastic dominance. We apply the new method to test the quantile treatment effect of the National Supported Work Demonstration program, which is a randomized employment enhancement program launched in the 1970s. The data associated with the participants of the program have been widely applied for treatment effect research since the pioneering study of LaLonde (1986). More recently, Delgado and Escanciano (2013) found that the program is beneficial for individuals of over 21 years of age. In our study, we find that the treatment tends to do better at raising the upper bounds of the earnings growth than raising the lower bounds. In other words, the program tends to increase the potential for high earnings growth but does not reduce the risk of negative earnings growth. The finding is particularly evident for those individuals who are older and spent more years at school. We should note that the tests based on the unconditional distribution cannot unveil the heterogeneity in the earnings growth quantiles in treatment effects.

The remaining part of this paper is organized as follows. In Section 3.2 we present our model, describe the estimators and state our asymptotic results. Section 3.3 is devoted to the bootstrap and we discuss its theoretical and practical aspects. The finite sample properties of both methods are investigated by means of a simulation study in Section 3.4, where we also compare the numerical performance of our method with the method proposed in Hall and Horowitz (2013) via simulations. The application of our new method is illustrated by a real data example in Section 3.5. The assumptions for our asymptotic theory are listed and discussed after the references. All detailed proofs are available in Appendix B.1.

3.2 Asymptotic confidence corridors

In Section 3.2.1 we present the prerequisites such as the precise definition of the model and a suitable estimate. The result on constructing confidence corridors (CCs) based on the distribution of the maximal absolute deviation are given in Section 3.2.2. In Section 3.2.3 we describe how to estimate the scaling factors, which appear in the limit theorems, using residual based estimators. Section 3.3.1 introduce a new bootstrap method for constructing CCs, while Section 3.3.2 is devoted to specific issues related to bootstrap CCs for quantile regression. Assumptions are listed and discussed after the references.

3.2.1 Prerequisites

Let $(\boldsymbol{X}_1, Y_1), ..., (\boldsymbol{X}_n, Y_n)$ be a sequence of independent identically distributed random vectors in $\mathbb{R}^{d+1}$ and consider the nonparametric regression model

$$Y_i = \theta_0(\boldsymbol{X}_i) + \varepsilon_i, \quad i = 1, ..., n, \tag{3.2.1}$$

where θ_0 is an aspect of Y conditional on $\boldsymbol{X}$, such as the τ-quantile, the τ-expectile or the mean regression curve, and the model errors $\varepsilon_1, \ldots, \varepsilon_n$ are i.i.d. with τ-

quantile, τ-expectile or mean equal to 0, respectively, depending on which θ_0 is in the model. The function $\theta(\boldsymbol{x})$ can be estimated by:

$$\hat{\theta}(\boldsymbol{x}) = \arg\min_{\theta\in\mathbb{R}} \frac{1}{n}\sum_{i=1}^{n} K_h(\boldsymbol{x}-\boldsymbol{X}_i)\rho(Y_i-\theta), \tag{3.2.2}$$

where $K_h(\boldsymbol{u}) = h^{-d}K(\boldsymbol{u}/h)$ for some kernel function $K : \mathbb{R}^d \to \mathbb{R}$, and a loss-function $\rho_\tau : \mathbb{R} \to \mathbb{R}$. In this paper we are concerned with the construction of uniform confidence corridors for quantile as well as expectile regression curves when the predictor is multivariate, that is, we focus on the loss functions

$$\rho_\tau(u) = \big|\mathbf{1}(u<0)-\tau\big||u|^k,$$

for $k = 1$ and 2 associated with quantile and expectile regression. We derive the asymptotic distribution of the properly scaled maximal deviation $\sup_{\boldsymbol{x}\in\mathcal{D}} |\hat{\theta}_n(\boldsymbol{x}) - \theta(\boldsymbol{x})|$ for both cases, where $\mathcal{D} \subset \mathbb{R}^d$ is a compact subset. We use strong approximations of the empirical process, concentration inequalities for general Gaussian random fields and results from extreme value theory. To be precise, we show that

$$\mathrm{P}\left[(2\delta\log n)^{1/2}\Big\{\sup_{\boldsymbol{x}\in\mathcal{D}}\big|r_n(\boldsymbol{x})\big[\hat{\theta}_n(\boldsymbol{x})-\theta_0(\boldsymbol{x})\big]\big|/\|K\|_2 - d_n\Big\} < a\right] \to \exp\big\{-2\exp(-a)\big\}, \tag{3.2.3}$$

as $n \to \infty$, where $r_n(\boldsymbol{x})$ is a scaling factor which depends on $\boldsymbol{x}$, n and the loss function under consideration.

3.2.2 Asymptotic results

In this section we present our main theoretical results on the distribution of the uniform maximal deviation of the quantile and expectile estimator. The proofs of the theorems at their full lengths are deferred to the appendix. Here we only give a brief sketch of proof of Theorem 3.2.1 which is the limit theorem for the case of quantile regression.

THEOREM 3.2.1. Let $\hat{\theta}_n(\boldsymbol{x})$ and $\theta_0(\boldsymbol{x})$ be the local constant quantile estimator and the true quantile function, respectively and suppose that assumptions (A1)-(A6) in Section B.1 hold. Let further $\mathrm{vol}(\mathcal{D}) = 1$ and

$$\begin{aligned} d_n =& (2d\kappa\log n)^{1/2} \\ &+ \big\{2d\kappa(\log n)\big\}^{-1/2}\left[\frac{1}{2}(d-1)\log\log n^\kappa + \log\big\{(2\pi)^{-1/2}H_2(2d)^{(d-1)/2}\big\}\right], \end{aligned}$$

where $H_2 = \big(2\pi\|K\|_2^2\big)^{-d/2}\det(\Sigma)^{1/2}$, $\Sigma = \big(\Sigma_{ij}\big)_{1\le i,j\le d} = \Big(\int \frac{\partial K(\boldsymbol{u})}{\partial u_i}\frac{\partial K(\boldsymbol{u})}{\partial u_j}d\boldsymbol{u}\Big)_{1\le i,j\le d}$,

$$r_n(\boldsymbol{x}) = \sqrt{\frac{nh^d f_{\boldsymbol{X}}(\boldsymbol{x})}{\tau(1-\tau)}}\, f_{Y|\boldsymbol{X}}\big\{\theta_0(\boldsymbol{x})|\boldsymbol{x}\big\},$$

Then the limit theorem (3.2.3) holds.

Sketch of proof. A major technical difficulty is imposed by the fact that the loss-function ρ_τ is not smooth which means that standard arguments such as those based on Taylor's theorem do not apply. As a consequence the use of a different, extended methodology becomes necessary. In this context Kong et al. (2010) derived a uniform Bahadur representation for an M-regression function in a multivariate setting (see appendix). It holds uniformly for $x \in \mathcal{D}$, where $\mathcal{D}$ is a compact subset of $\mathbb{R}^d$:

$$\hat{\theta}_n(\boldsymbol{x}) - \theta_0(\boldsymbol{x}) = \frac{1}{nS_{n,0,0}(\boldsymbol{x})} \sum_{i=1}^{n} K_h(\boldsymbol{x} - \boldsymbol{X}_i)\psi_\tau\big\{Y_i - \theta_0(\boldsymbol{x})\big\} + \mathcal{O}\left\{\Big(\frac{\log n}{nh^d}\Big)^{\frac{3}{4}}\right\}, \ a.s. \tag{3.2.4}$$

Here $S_{n,0,0}(\boldsymbol{x}) = \int K(\boldsymbol{u})g(\boldsymbol{x}+h\boldsymbol{u})f_{\boldsymbol{X}}(\boldsymbol{x}+h\boldsymbol{u})d\boldsymbol{u}$, $\psi_\tau(u) = \mathbf{1}(u < 0) - \tau$ is the piecewise derivative of the loss-function ρ_τ and

$$g(\boldsymbol{x}) = \frac{\partial}{\partial t}\mathsf{E}[\psi_\tau(Y-t)|\boldsymbol{X} = \boldsymbol{x}]\Big|_{t=\theta_0(\boldsymbol{x})}.$$

Notice that the error term of the Bahadur expansion does not depend on the design $\boldsymbol{X}$ and it converges to 0 with rate $\big(\log n/nh^d\big)^{\frac{3}{4}}$ which is much faster than the convergence rate $(nh^d)^{-\frac{1}{2}}$ of the stochastic term.

Rearranging (3.2.4), we obtain

$$S_{n,0,0}(\boldsymbol{x})\{\hat{\theta}_n(\boldsymbol{x}) - \theta_0(\boldsymbol{x})\} = \frac{1}{n}\sum_{i=1}^{n} K_h(\boldsymbol{x} - \boldsymbol{X}_i)\psi_\tau\big\{Y_i - \theta_0(\boldsymbol{x})\big\} + \mathcal{O}\left\{\Big(\frac{\log n}{nh^d}\Big)^{\frac{3}{4}}\right\}. \tag{3.2.5}$$

Now we express the leading term on the right hand side of (3.2.5) by means of the centered empirical process

$$Z_n(y, \boldsymbol{u}) = n^{1/2}\{F_n(y, \boldsymbol{u}) - F(y, \boldsymbol{u})\}, \tag{3.2.6}$$

where $F_n(y, \boldsymbol{x}) = n^{-1}\sum_{i=1}^{n} \mathbf{1}(Y_i \le y, X_{i1} \le x_1, ..., X_{id} \le x_d)$. This yields, by Fubini's theorem,

$$\begin{aligned} &S_{n,0,0}(\boldsymbol{x})\{\hat{\theta}_n(\boldsymbol{x}) - \theta_0(\boldsymbol{x})\} - b(\boldsymbol{x}) \\ &= n^{-1/2}\int\int K_h(\boldsymbol{x}-\boldsymbol{u})\psi_\tau\big\{y - \theta_0(\boldsymbol{x})\big\}dZ_n(y, \boldsymbol{u}) + \mathcal{O}\left\{\Big(\frac{\log n}{nh^d}\Big)^{\frac{3}{4}}\right\}, \end{aligned} \tag{3.2.7}$$

where

$$b(\boldsymbol{x}) = -\mathsf{E}_{\boldsymbol{x}}\left[\frac{1}{n}\sum_{i=1}^{n} K_h(\boldsymbol{x} - \boldsymbol{X}_i)\psi\big\{Y_i - \theta_0(\boldsymbol{x})\big\}\right]$$

denotes the bias which is of order $\mathcal{O}(h^s)$ by Assumption (A3) in the Appendix. The variance of the first term of the right hand side of (3.2.7) can be estimated via a

change of variables and Assumption (A5), which gives

$$
\begin{aligned}
&(nh^d)^{-2} n \mathsf{E}\big[K^2\big\{(\boldsymbol{x}-\boldsymbol{X}_i)/h\big\}\psi^2\big\{Y_i-\theta_0(\boldsymbol{x})\big\}\big] \\
&= (nh^d)^{-2} nh^d \int\int K^2(\boldsymbol{v})\psi^2\big\{y-\theta_0(\boldsymbol{x})\big\} f_{Y|\boldsymbol{X}}(y|\boldsymbol{x}-h\boldsymbol{v}) f_{\boldsymbol{X}}(\boldsymbol{x}-h\boldsymbol{v}) dy d\boldsymbol{v} \\
&= (nh^d)^{-1} \int\int K^2(\boldsymbol{v})\psi^2\big\{y-\theta_0(\boldsymbol{x})\big\} f_{Y|\boldsymbol{X}}(y|\boldsymbol{x}) f_{\boldsymbol{X}}(\boldsymbol{x}) dy d\boldsymbol{v} + \mathcal{O}\big((nh^{d-1})^{-1}\big) \\
&= (nh^d)^{-1} f_{\boldsymbol{X}}(\boldsymbol{x})\sigma^2(\boldsymbol{x})\|K\|_2^2 + \mathcal{O}\big\{(nh^d)^{-1}h\big\},
\end{aligned}
$$

where $\sigma^2(\boldsymbol{x}) = \mathsf{E}[\psi^2\big\{Y-\theta_0(\boldsymbol{x})\big\}|\boldsymbol{X}=\boldsymbol{x}]$. The standardized version of (3.2.5) can therefore be approximated by

$$
\begin{aligned}
&\frac{\sqrt{nh^d}}{\sqrt{f_{\boldsymbol{X}}(\boldsymbol{x})}\sigma(\boldsymbol{x})\|K\|_2} S_{n,0,0}(\boldsymbol{x})\{\hat{\theta}_n(\boldsymbol{x})-\theta_0(\boldsymbol{x})\} \\
&= \frac{1}{\sqrt{h^d f_{\boldsymbol{X}}(\boldsymbol{x})}\sigma(\boldsymbol{x})\|K\|_2} \int\int K\left(\frac{\boldsymbol{x}-\boldsymbol{u}}{h}\right)\psi\big\{Y_i-\theta_0(\boldsymbol{x})\big\} dZ_n(y,\boldsymbol{u}) \\
&\qquad\qquad + \mathcal{O}\big(\sqrt{nh^d}h^s\big) + \mathcal{O}\left\{\left(\frac{\log n}{nh^d}\right)^{\frac{3}{4}}\right\}. \qquad (3.2.8)
\end{aligned}
$$

The dominating term is defined by

$$
Y_n(\boldsymbol{x}) \stackrel{\text{def}}{=} \frac{1}{\sqrt{h^d f_{\boldsymbol{X}}(\boldsymbol{x})}\sigma(\boldsymbol{x})} \int\int K\left(\frac{\boldsymbol{x}-\boldsymbol{u}}{h}\right)\psi\big\{y-\theta_0(\boldsymbol{x})\big\} dZ_n(y,\boldsymbol{u}). \qquad (3.2.9)
$$

Involving strong Gaussian approximation and Bernstein-type concentration inequalities, this process can be approximated by a stationary Gaussian field:

$$
Y_{5,n}(\boldsymbol{x}) = \frac{1}{\sqrt{h^d}} \int K\left(\frac{\boldsymbol{x}-\boldsymbol{u}}{h}\right) dW\big(\boldsymbol{u}\big), \qquad (3.2.10)
$$

where W denotes a Brownian sheet. The supremum of this process is asymptotically Gumbel distributed, which follows, e.g., by Theorem 2 of Rosenblatt (1976). Since the kernel is symmetric and of order s, we can estimate the term

$$
S_{n,0,0} = f_{Y|\boldsymbol{X}}(\theta_0(\boldsymbol{x})|\boldsymbol{x}) f_{\boldsymbol{X}}(\boldsymbol{x}) + \mathcal{O}(h^s)
$$

if (A5) holds. On the other hand, $\sigma^2(\boldsymbol{x}) = \tau(1-\tau)$ in quantile regression. Therefore, the statements of the theorem hold.

□

Detailed proof of Theorem 3.2.1 can be found in Appendix B.1.1.

COROLLARY 3.2.2 (CC for multivariate quantile regression)**.** Under the assumptions of Theorem 3.2.1, an approximate $(1-\alpha)\times 100\%$ confidence corridor is given by

$$
\hat{\theta}_n(\boldsymbol{t}) \pm (nh^d)^{-1/2}\big\{\tau(1-\tau)\|K\|_2/\hat{f}_{\boldsymbol{X}}(\boldsymbol{t})\big\}^{1/2}\hat{f}_{\varepsilon|\boldsymbol{X}}\big\{0|\boldsymbol{t}\big\}^{-1}\Big\{d_n + c(\alpha)(2\kappa d\log n)^{-1/2}\Big\},
$$

where $\alpha \in (0,1)$ and $c(\alpha) = \log 2 - \log\big|\log(1-\alpha)\big|$ and $\hat{f}_{\boldsymbol{X}}(\boldsymbol{t})$, $\hat{f}_{\varepsilon|\boldsymbol{X}}\big\{0|\boldsymbol{t}\big\}$ are consistent estimates for $f_{\boldsymbol{X}}(\boldsymbol{t})$, $f_{\varepsilon|\boldsymbol{X}}\big\{0|\boldsymbol{t}\big\}$ with convergence rate in sup-norm faster than $o_p\big((\log n)^{-1/2}\big)$.

REMARK 3.2.3. Note that under the conditions of Corollary 3.2.2 we find

$$\sup_{\boldsymbol{x}\in\mathcal{D}} \big|r_n(\boldsymbol{x})\big(\hat{\theta}_n(\boldsymbol{x}) - \theta_0(\boldsymbol{x}))\big)\big| = \mathcal{O}_P\big(\sqrt{\log(n)}\big),$$

where

$$r_n(\boldsymbol{x}) = \sqrt{\frac{nh^d f_{\boldsymbol{X}}(\boldsymbol{x})}{\tau(1-\tau)}} f_{Y|\boldsymbol{X}}\big\{\theta_0(\boldsymbol{x})|\boldsymbol{x}\big\}.$$

For kernel estimators $\hat{f}_{\varepsilon|\boldsymbol{X}}(0,\cdot)$ and $\hat{f}_{\boldsymbol{X}}(\cdot)$ converging in sup-norm with error rate $o_P\big\{(\log n)^{-1/2}\big\}$ to $f_{\varepsilon|\boldsymbol{X}}(0,\cdot)$ and $f_{\boldsymbol{X}}(\cdot)$, respectively, the quantity $\hat{r}_n(\boldsymbol{x})$, defined by

$$\hat{r}_n(\boldsymbol{x}) = \sqrt{\frac{nh^d \hat{f}_{\boldsymbol{X}}(\boldsymbol{x})}{\tau(1-\tau)}} \hat{f}_{\varepsilon|\boldsymbol{X}}(0,\boldsymbol{x}),$$

inherits this rate. Furthermore, since we consider an additive error model, the conditional density $f_{Y|\boldsymbol{X}}\big\{\theta_0(\boldsymbol{x})|\boldsymbol{x}\big\}$ can be replaced by $f_{\varepsilon|\boldsymbol{X}}(0,\boldsymbol{x})$ (see Section 3.2.3 below for more details and the definition of suitable estimators). This yields

$$\sup_{\boldsymbol{x}\in\mathcal{D}} \big|\hat{r}_n(\boldsymbol{x})\big(\hat{\theta}_n(\boldsymbol{x}) - \theta_0(\boldsymbol{x}))\big)\big| = o_P(1) + \sup_{\boldsymbol{x}\in\mathcal{D}} \big|r_n(\boldsymbol{x})\big(\hat{\theta}_n(\boldsymbol{x}) - \theta_0(\boldsymbol{x}))\big)\big|.$$

Hence, by Slutsky's Lemma, the quantities $\sup_{\boldsymbol{x}\in\mathcal{D}} \big|\hat{r}_n(\boldsymbol{x})\big(\hat{\theta}_n(\boldsymbol{x}) - \theta_0(\boldsymbol{x}))\big)\big|$ and $\sup_{\boldsymbol{x}\in\mathcal{D}} \big|r_n(\boldsymbol{x})\big(\hat{\theta}_n(\boldsymbol{x}) - \theta_0(\boldsymbol{x}))\big)\big|$ have the same asymptotic distribution.

The expectile confidence corridor can be constructed in an analogous manner as the quantile confidence corridor. The two cases differ in the form and hence the properties of the loss function. Therefore we find for expectile regression:

$$S_{n,0,0}(\boldsymbol{x}) = -2\big[F_{Y|\boldsymbol{X}}\big(\theta_0(\boldsymbol{x})\big)|\boldsymbol{x})(2\tau-1) - \tau\big] f_{\boldsymbol{X}}(\boldsymbol{x}) + \mathcal{O}(h^s).$$

Through similar approximation steps as the quantile regression, we derive the following theorem.

THEOREM 3.2.4. Let $\hat{\theta}_n(\boldsymbol{x})$ be the the local constant expectile estimator and $\theta_0(\boldsymbol{x})$ the true expectile function. If Assumptions (A1), (A3)-(A6) and (EA2) of Section B.1 hold with a constant b_1 satisfying

$$n^{-1/6} h^{-d/2-3d/(b_1-2)} = \mathcal{O}(n^{-\nu}), \quad \nu > 0.$$

Then the limit theorem (3.2.3) holds with a scaling factor

$$r_n(\boldsymbol{x}) = \sqrt{nh^d f_{\boldsymbol{X}}(\boldsymbol{x})}\sigma^{-1}(\boldsymbol{x}) \left\{2\big[\tau - F_{Y|\boldsymbol{X}}(\theta_0(\boldsymbol{x})|\boldsymbol{x})(2\tau-1)\big]\right\}$$

and with the same constants H_2 and d_n as defined in Theorem 3.2.1, where $\sigma^2(\boldsymbol{x}) = \mathsf{E}[\psi_\tau^2(Y-\theta_0(\boldsymbol{x}))|\boldsymbol{X}=\boldsymbol{x}]$ and $\psi_\tau(u) = 2(\mathbf{1}(u\le 0)-\tau)|u|$ is the derivative of the expectile loss-function $\rho_\tau(u) = \big|\tau - \mathbf{1}(u<0)\big||u|^2$.

The proof of this result is deferred to Appendix B.1.2. In the next corollary, the explicit form of the CCs for expectiles is given.

COROLLARY 3.2.5 (CC for multivariate expectile regression)**.** Under the same assumptions of Theorem 3.2.4, an approximate $(1-\alpha)\times 100\%$ confidence corridor is given by

$$\hat{\theta}_n(\boldsymbol{t}) \pm (nh^d)^{-1/2}\big\{\hat{\sigma}^2(\boldsymbol{t})\|K\|_2/\hat{f}_{\boldsymbol{X}}(\boldsymbol{t})\big\}^{1/2} \\ \Big\{-2\big[\hat{F}_{\varepsilon|\boldsymbol{X}}\big\{0|\boldsymbol{t}\big\}(2\tau-1)-\tau\big]\Big\}^{-1}\Big\{d_n + c(\alpha)(2\kappa d\log n)^{-1/2}\Big\},$$

where $\alpha \in (0,1)$ $c(\alpha) = \log 2 - \log\big|\log(1-\alpha)\big|$ and $\hat{f}_{\boldsymbol{X}}(\boldsymbol{t})$, $\hat{\sigma}^2(\boldsymbol{t})$ and $\hat{F}_{\varepsilon|\boldsymbol{X}}(0|\boldsymbol{x})$ are consistent estimates for $f_{\boldsymbol{X}}(\boldsymbol{t})$, $\sigma^2(\boldsymbol{t})$ and $F_{\varepsilon|\boldsymbol{X}}(0|\boldsymbol{x})$ with convergence rate in sup-norm faster than $\mathcal{o}_p\big((\log n)^{-1/2}\big)$.

A further immediate consequence of Theorem 3.2.4 is a similar limit theorem in the context of local least squares estimation of the regression curve in classical mean regression.

COROLLARY 3.2.6 (CC for multivariate mean regression)**.** Consider the loss function $\rho(u) = u^2$ corresponding to $\psi(u) = 2u$. Under the assumptions of Theorem 3.2.4, with the same constants H_2 and d_n, (3.2.3) holds for the local constant estimator $\hat{\theta}$ and the regression function $\theta(x) = E[Y \mid \boldsymbol{X} = \boldsymbol{x}]$ with scaling factor $r(\boldsymbol{x}) = \sqrt{nh^d f_{\boldsymbol{X}}(\boldsymbol{x})}\sigma^{-1}(\boldsymbol{x})$ and $\sigma^2(\boldsymbol{x}) = \mathrm{Var}[Y \mid \boldsymbol{X} = \boldsymbol{x}]$.

REMARK 3.2.7. We would like to stress that our purely non-parametric approach offers flexibility and reasonable results in moderate dimensions $d = 2$, $d = 3$, but it is not suitable for inference in high dimensional models due to the curse of dimensionality. The case of high dimensional regressors may be handled via a semi-parametric specification of the regression curve, such as, for instance, a partial linear model. Such a model was considered in Song et al. (2012) with a one-dimensional non-parametric component. We think that our approach allows to adapt these ideas and, as an extension, to consider a non-parametric component which is multivariate. Hence, our approach then also offers higher flexibility in semi-parametric modeling. This semi-parametric approach is not pursued further in this paper but it clearly deserves future research.

3.2.3 Estimating the scaling factors

The performance of the confidence bands is greatly influenced by the scaling factors $\hat{f}_{\varepsilon|\boldsymbol{X}}(v|\boldsymbol{x})$, $F_{\varepsilon|\boldsymbol{X}}(v|\boldsymbol{x})$ and $\hat{\sigma}(\boldsymbol{x})^2$. The purpose of this subsection is thus to propose a way to estimate these factors and investigate their asymptotic properties.

As pointed out by our referee, estimating $f_{\varepsilon|\boldsymbol{X}}(0)$ is not a trivial task. The application of a rank test described in Chapter 3.5 of Koenker (2005) is an alternative to avoid estimating $f_{\varepsilon|\boldsymbol{X}}(0)$ in parametric quantile regression. However, it is a challenging task to apply this technique to kernel smoothing quantile regression. For

pointwise nonparametric inference, it may be possible to construct a test by adding weights (given by $h^{-1}K((\boldsymbol{x}-\boldsymbol{X}_i)/h)$, where h is the bandwidth and K is the kernel function) in the linear programing problem and therefore its dual can also be computed. However, a global shape test like the one investigated in this paper cannot be derived from the rank test. Hence, it seems inevitable to estimate the nuisance parameters and plug them into the test statistics.

Since we consider the additive error model (3.2.1), the conditional distribution function $F_{Y|\boldsymbol{X}}(\theta_0(\boldsymbol{x})|\boldsymbol{x})$ and the conditional density $f_{Y|\boldsymbol{X}}(\theta_0(\boldsymbol{x})|\boldsymbol{x})$ can be replaced by $F_{\varepsilon|\boldsymbol{X}}(0|\boldsymbol{x})$ and $f_{\varepsilon|\boldsymbol{X}}(0|\boldsymbol{x})$, respectively, where $F_{\varepsilon|\boldsymbol{X}}$ and $f_{\varepsilon|\boldsymbol{X}}$ are the conditional distribution and density functions of ε. Similarly, we have

$$\sigma^2(\boldsymbol{x}) = \mathsf{E}\big[\psi_\tau\big(Y-\theta_0(\boldsymbol{x})\big)^2\big|\boldsymbol{X}=\boldsymbol{x}\big] = \mathsf{E}\big[\psi_\tau(\varepsilon)^2\big|\boldsymbol{X}=\boldsymbol{x}\big]$$

where ε may depend on $\boldsymbol{X}$ due to heterogeneity. It should be noted that the kernel estimators for $f_{\varepsilon|\boldsymbol{X}}(0|\boldsymbol{x})$ and $f_{Y|\boldsymbol{X}}(\theta_0(\boldsymbol{x})|\boldsymbol{x})$ are asymptotically equivalent, but show different finite sample behavior. We explore this issue further in the following section.

Introducing the residuals $\hat{\varepsilon}_i = Y_i - \hat{\theta}_n(\boldsymbol{X}_i)$ we propose to estimate $F_{\varepsilon|\boldsymbol{X}}$, $f_{\varepsilon|\boldsymbol{X}}$ and $\sigma^2(\boldsymbol{x})$ by

$$\hat{F}_{\varepsilon|\boldsymbol{X}}(v|\boldsymbol{x}) = n^{-1}\sum_{i=1}^{n} G\left(\frac{v-\hat{\varepsilon}_i}{h_0}\right) L_{\bar{h}}(\boldsymbol{x}-\boldsymbol{X}_i)/\hat{f}_{\boldsymbol{X}}(\boldsymbol{x}), \tag{3.2.11}$$

$$\hat{f}_{\varepsilon|\boldsymbol{X}}(v|\boldsymbol{x}) = n^{-1}\sum_{i=1}^{n} g_{h_0}\left(v-\hat{\varepsilon}_i\right) L_{\bar{h}}(\boldsymbol{x}-\boldsymbol{X}_i)/\hat{f}_{\boldsymbol{X}}(\boldsymbol{x}), \tag{3.2.12}$$

$$\hat{\sigma}^2(\boldsymbol{x}) = n^{-1}\sum_{i=1}^{n} \psi^2(\hat{\varepsilon}_i) L_{\bar{h}}(\boldsymbol{x}-\boldsymbol{X}_i)/\hat{f}_{\boldsymbol{X}}(\boldsymbol{x}), \tag{3.2.13}$$

where $\hat{f}_{\boldsymbol{X}}(\boldsymbol{x}) = n^{-1}\sum_{i=1}^{n} L_{\bar{h}}(\boldsymbol{x}-\boldsymbol{X}_i)$, G is a given continuously differentiable cumulative distribution function and g is its derivative. The construction of estimators in (3.2.11) and (3.2.12) follows from the estimator for general conditional distribution and density functions discussed in Chapter 5 and 6 of Li and Racine (2007). The same bandwidth $\bar{h}$ is applied to the three estimators, but the choice of $\bar{h}$ will make the convergence rate of (3.2.13) sub-optimal. More details on the choice of $\bar{h}$ are given in section 3.3.2 below. Nevertheless, the rate of convergence of (3.2.13) is of polynomial order in n. The theory developed in this subsection can be generalized to the case of different bandwidth for different direction without much difficulty.

The estimators (3.2.11) and (3.2.12) belong to the family of residual-based estimators. The consistency of residual-based density estimators for errors in a regression model are explored in the literature in various settings. It is possible to obtain an expression for the residual based kernel density estimator as the sum of the estimator with the true residuals, the partial sum of the true residuals and a term for the bias of the nonparametrically estimated function, as shown in Muhsal and Neumeyer (2010), among others. The residual based conditional kernel density case is less considered in the literature. Kiwitt and Neumeyer (2012) consider the

residual based kernel estimator for conditional distribution function conditioning on a one-dimensional variable.

Below we give consistency results for the estimators defined in (3.2.11), (3.2.12) and (3.2.13). The proof can be found in the appendix.

LEMMA 3.2.8. Under conditions (A1), (A3)-(A5), (B1)-(B3) in Section B.1, we have

1) $\sup_{v\in I}\sup_{\boldsymbol{x}\in\mathcal{D}}\left|\hat{F}_{\varepsilon|\boldsymbol{X}}(v|\boldsymbol{x})-F_{\varepsilon|\boldsymbol{X}}(v|\boldsymbol{x})\right|=\mathcal{O}_p\big(t_n\big)$,

2) $\sup_{v\in I}\sup_{\boldsymbol{x}\in\mathcal{D}}\left|\hat{f}_{\varepsilon|\boldsymbol{X}}(v|\boldsymbol{x})-f_{\varepsilon|\boldsymbol{X}}(v|\boldsymbol{x})\right|=\mathcal{O}_p\big(t_n\big)$,

3) $\sup_{\boldsymbol{x}\in\mathcal{D}}\left|\hat{\sigma}^2(\boldsymbol{x})-\sigma^2(\boldsymbol{x})\right|=\mathcal{O}_p\big(u_n\big)$,

where $t_n=\mathcal{O}\big\{h_0^{s'}+h^s+\bar{h}^{s'}+(n\bar{h}^d)^{-1/2}\log n+(nh^d)^{-1/2}\log n\big\}=\mathcal{O}(n^{-\lambda})$, and $u_n=\mathcal{O}\big\{h^s+\bar{h}^{s'}+(n\bar{h}^d)^{-1/2}\log n+(nh^d)^{-1/2}\log n\big\}=\mathcal{O}(n^{-\lambda_1})$ for some constants $\lambda,\lambda_1>0$.

Detailed proof of Lemma 3.2.8 can be found in Appendix B.1.3. The factor of $\log n$ shown in the convergence rate is the price which we pay for the supnorm deviation. Since these estimators uniformly converge in a polynomial rate in n, the asymptotic distributions in Theorem 3.2.1 and 3.2.4 do not change if we plug these estimators into the formulae.

3.3 Bootstrap confidence corridors

3.3.1 Asymptotic theory

In the case of the suitably normed maximum of independent standard normal variables, it is shown in Hall (1979) that the speed of convergence in limit theorems of the form (3.2.3) is of order $1/\log n$, that is, the coverage error of the asymptotic CC decays only logarithmically. This leads to unsatisfactory finite sample performance of the asymptotic methods, especially for small sample sizes and dimensions $d>1$. However, Hall (1991) suggests that the use of a bootstrap method, based on a proper way of resampling, can increase the speed of shrinking of coverage error to a polynomial rate of n. In this section we therefore propose a specific bootstrap technique and construct a confidence corridor for the objects to be analysed.

Given the residuals $\hat{\varepsilon}_i=Y_i-\hat{\theta}_n(\boldsymbol{X}_i)$, the bootstrap observations $(\boldsymbol{X}_i^*,\varepsilon_i^*)$ are sampled from

$$\hat{f}_{\varepsilon,\boldsymbol{X}}(v,\boldsymbol{x})=\frac{1}{n}\sum_{i=1}^{n}g_{h_0}\left(\hat{\varepsilon}_i-v\right)L_{\bar{h}}(\boldsymbol{x}-\boldsymbol{X}_i),\tag{3.3.1}$$

where g and L are a kernel functions with bandwidths h_0, $\bar{h}$ satisfying assumptions (B1)-(B3). In particular, in our simulation study, we choose L to be a product

Gaussian kernel. In the following discussion P^* and E^* stand for the probability and expectation conditional on the data $(\boldsymbol{X}_i, Y_i)$, $i = 1, ..., n$.

We introduce the notation

$$A_n^*(\boldsymbol{x}) = \frac{1}{n}\sum_{i=1}^{n} K_h(\boldsymbol{x} - \boldsymbol{X}_i^*)\psi_\tau(\varepsilon_i^*),$$

and define the so-called "one-step estimator" $\theta^*(\boldsymbol{x})$ from the bootstrap sample by

$$\hat{\theta}^*(\boldsymbol{x}) - \hat{\theta}_n(\boldsymbol{x}) = \hat{S}_{n,0,0}^{-1}(\boldsymbol{x})\left\{A_n^*(\boldsymbol{x}) - \mathsf{E}^*[A_n^*(\boldsymbol{x})]\right\}, \tag{3.3.2}$$

where

$$\hat{S}_{n,0,0}(\boldsymbol{x}) = \begin{cases} \hat{f}_{\varepsilon|\boldsymbol{X}}\big(0|\boldsymbol{x}\big)\hat{f}_{\boldsymbol{X}}(\boldsymbol{x}), & \text{quantile case;} \\ 2\big\{\tau - \hat{F}_{\varepsilon|\boldsymbol{X}}\big(0|\boldsymbol{x}\big)(2\tau - 1)\big\}\hat{f}_{\boldsymbol{X}}(\boldsymbol{x}), & \text{expectile case.} \end{cases} \tag{3.3.3}$$

note that $\mathsf{E}^*[\hat{\theta}^*(\boldsymbol{x}) - \hat{\theta}_n(\boldsymbol{x})] = 0$, so $\hat{\theta}^*(\boldsymbol{x})$ is unbiased for $\hat{\theta}_n(\boldsymbol{x})$ under E^*. As a remark, we note that undersmoothing is applied in our procedure for two reasons: first, the theory we developed so far is based on undersmoothing; secondly, it is suggested in Hall (1992) that undersmoothing is more effective than oversmoothing given that the goal is to achieve coverage accuracy.

Note that the bootstrap estimate (3.3.2) is motivated by the smoothed bootstrap procedure proposed in Claeskens and Van Keilegom (2003). In contrast to these authors we make use of the leading term of the Bahadur representation. Mammen et al. (2013) also use the leading term of a Bahadur representation proposed in Guerre and Sabbah (2012) to construct bootstrap samples. Song et al. (2012) propose a bootstrap for quantile regression based on oversmoothing, which has the drawback that it requires iterative estimation, and oversmoothing is in general less effective in terms of coverage accuracy.

For the following discussion define

$$Y_n^*(\boldsymbol{x}) = \frac{1}{\sqrt{h^d \hat{f}_{\boldsymbol{X}}(\boldsymbol{x})}\sigma_*(\boldsymbol{x})} \int\int K\left(\frac{\boldsymbol{x} - \boldsymbol{u}}{h}\right)\psi_\tau\big(v\big)dZ_n^*(v, \boldsymbol{u}) \tag{3.3.4}$$

as the bootstrap analogue of the process (3.2.9), where

$$Z_n^*(y, \boldsymbol{u}) = n^{1/2}\left\{F_n^*(v, \boldsymbol{u}) - \hat{F}(v, \boldsymbol{u})\right\}, \quad \sigma_*(\boldsymbol{x}) = \sqrt{\mathsf{E}^*\left[\psi_\tau(\varepsilon_i^*)^2|\boldsymbol{x}\right]} \tag{3.3.5}$$

and

$$F_n^*(v, \boldsymbol{u}) = \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}\left\{\varepsilon_i^* \le v, X_1^* \le u_1, ..., X_d^* \le u_d\right\}.$$

The process Y_n^* serves as an approximation of a standardized version of $\hat{\theta}_n^* - \hat{\theta}_n$, and similar to the previous sections the process Y_n^* is approximated by a stationary Gaussian field $Y_{n,5}^*$ under P^* with probability one, that is,

$$Y_{5,n}^*(\boldsymbol{x}) = \frac{1}{\sqrt{h^d}} \int K\left(\frac{\boldsymbol{x} - \boldsymbol{u}}{h}\right) dW^*(\boldsymbol{u}).$$

Finally, $\sup_{\boldsymbol{x}\in\mathcal{D}}\left|Y^*_{5,n}(\boldsymbol{x})\right|$ is asymptotically Gumbel distributed conditional on samples.

THEOREM 3.3.1. Suppose that assumptions (A1)-(A6), (C1) in Section B.1 hold, and $\mathrm{vol}(\mathcal{D}) = 1$, let

$$r^*_n(\boldsymbol{x}) = \sqrt{\frac{nh^d}{\hat{f}_{\boldsymbol{X}}(\boldsymbol{x})\sigma^2_*(\boldsymbol{x})}}\hat{S}_{n,0,0}(\boldsymbol{x}),$$

where $\hat{S}_{n,0,0}(\boldsymbol{x})$ is defined in (3.3.3) and $\sigma^2_*(\boldsymbol{x})$ is defined in (3.3.5). Then

$$\mathrm{P}^*\left\{(2d\kappa\log n)^{1/2}\left(\sup_{\boldsymbol{x}\in\mathcal{D}}\left[r^*_n(\boldsymbol{x})|\hat{\theta}^*(\boldsymbol{x}) - \hat{\theta}_n(\boldsymbol{x})|\right]/\|K\|_2 - d_n\right) < a\right\} \\ \to \exp\left\{-2\exp(-a)\right\},\ \text{a.s.} \qquad (3.3.6)$$

as $n\to\infty$ for the local constant quantile regression estimate. If (A1)-(A6) and (EC1) hold with a constant $b \geq 4$ satisfying

$$n^{-\frac{1}{6}+\frac{4}{b^2}-\frac{1}{b}}h^{-\frac{d}{2}-\frac{6d}{b}} = \mathcal{O}(n^{-\nu}), \quad \nu > 0,$$

then (3.3.6) also holds for expectile regression with corresponding $\sigma^2_*(\boldsymbol{x})$.

The proof can be found in Appendix B.1.4. The following lemma suggests that we can replace $\sigma_*(\boldsymbol{x})$ in the limiting theorem by $\hat{\sigma}(\boldsymbol{x})$.

LEMMA 3.3.2. If assumptions (B1)-(B3), and (EC1) in Section B.1 are satisfied with $b > 2(2s' + d + 1)/(2s' + 3)$, then

$$\|\sigma^2_*(\boldsymbol{x}) - \hat{\sigma}^2(\boldsymbol{x})\| = o^*_p\big((\log n)^{-1/2}\big), \quad a.s.$$

The following corollary is a consequence of Theorem 3.3.1.

COROLLARY 3.3.3. Under the same conditions as stated in Theorem 3.3.1, the (asymptotic) bootstrap confidence set of level $1-\alpha$ is given by

$$\left\{\theta : \sup_{\boldsymbol{x}\in\mathcal{D}}\left|\frac{\hat{S}_{n,0,0}(\boldsymbol{x})}{\sqrt{\hat{f}_{\boldsymbol{X}}(\boldsymbol{x})\hat{\sigma}^2(\boldsymbol{x})}}\big[\hat{\theta}_n(\boldsymbol{x}) - \theta(\boldsymbol{x})\big]\right| \leq \xi^*_\alpha\right\}, \qquad (3.3.7)$$

where ξ^*_α satisfies

$$\lim_{n\to\infty}\mathrm{P}^*\left(\sup_{\boldsymbol{x}\in\mathcal{D}}\left|\frac{\hat{S}_{n,0,0}(\boldsymbol{x})}{\sqrt{\hat{f}_{\boldsymbol{X}}(\boldsymbol{x})\hat{\sigma}^2(\boldsymbol{x})}}\big[\hat{\theta}^*(\boldsymbol{x}) - \hat{\theta}_n(\boldsymbol{x})\big]\right| \leq \xi^*_\alpha\right) = 1-\alpha, \quad a.s. \qquad (3.3.8)$$

where $\hat{S}_{n,0,0}$ is defined in (3.3.3).

Note that it does not create much difference to standardize the $\hat{\theta}_n(\boldsymbol{x}) - \theta_0(\boldsymbol{x})$ in (3.3.6) with $\hat{f}_{\boldsymbol{X}}$ and $\hat{\sigma}^2(\boldsymbol{x})$ constructed from original samples or $\hat{f}_{\boldsymbol{X}}$ and $\hat{\sigma}^2(\boldsymbol{x})$ from the bootstrap samples. The simulation results of Claeskens and Van Keilegom (2003) show that the two ways of standardization give similar coverage probabilities for confidence corridors of kernel ML estimators.

3.3.2 Implementation

In this section, we discuss issues related to the implementation of the bootstrap for quantile regression.

Note that the *width* of the CC is determined by the variance and the *location* is affected by the bias of the quantile function estimator, and both depend on the bandwidth used for estimation. Hence, the choice of bandwidth needs to balance the bias (location) and the variance (size). It is chosen such that the bias is only just negligible after normalization, that is, slightly smaller than the L^2-optimal bandwidth. Therefore, it is enough to take an undersmoothed $h = \mathcal{O}(n^{-1/(2s+d)-\delta})$, given that $s > d$ and $\delta > 0$, where s is the order of Hölder continuity of the function θ_0 and δ is the degree of undersmoothing. We may use the methods proposed by Yu and Jones (1998) for nonparametric quantile regression to choose the bandwidth before undersmoothing, namely

$$h_{\tau,j} = h_{1,j}\{\tau(1-\tau)/\phi(\Phi^{-1}(\tau))^2\}^{1/5}, \quad j = 1, 2, \tag{3.3.9}$$

where $h_{1,j}$ are chosen by common methods like the rule-of-thumb or cross-validation for mean regression or density estimation and Φ is the CDF of the standard Gaussian distribution. In our simulation study, we select $h_{1,j}$ in (3.3.9) by the rule-of-thumb, implemented with the `np` package in `R`. In our application analysis, $h_{1,j}$ in (3.3.9) are chosen by the cross-validated bandwidth for the conditional distribution smoother of Y given $\boldsymbol{X}$, implemented with the `np` package in `R`. This package is based on the paper of Li et al. (2013).

For expectile regression, we use the rule-of-thumb bandwidth for the conditional distribution smoother of Y given $\boldsymbol{X}$, chosen with the `np` package in `R`.

The choice of h_0 and $\bar{h}$ for estimating the scaling factors in Section 3.2.3 should minimize the convergence rate of these residual based estimators. Hence, observing that the terms related to h_0 and $\bar{h}$ are similar to those in usual $(d+1)$-dimensional density estimators, it is reasonable to choose $h_0 \sim \bar{h} \sim n^{-1/(5+d)}$, given that L, g are second order kernels. We choose the rule-of-thumb bandwidths for conditional densities with the `R` package `np` in our simulation and application studies.

The one-step estimator for quantile regression defined in (3.3.2) depends sensitively on the estimator of $\hat{S}_{n,0,0}(\boldsymbol{x})$. Unlike in the expectile case, the function $\psi(\cdot)$ in the quantile case is bounded, and, as a result, the bootstrapped density based on (3.3.7) is very easily influenced by the factor $\hat{S}_{n,0,0}(\boldsymbol{x})$; in particular, $\hat{f}_{\varepsilon|\boldsymbol{X}}(0|\boldsymbol{x})$. As pointed out by Feng et al. (2011), the residual of quantile regression tends to be less dispersed than the model error; thus $\hat{f}_{\varepsilon|\boldsymbol{X}}(0|\boldsymbol{x})$ tends to over-estimate the true $f_{\varepsilon|\boldsymbol{X}}(0|\boldsymbol{x})$ for each $\boldsymbol{x}$.

The way of getting around this problem is based on the following observation: An additive error model implies the equality $f_{Y|\boldsymbol{X}}\big\{v + \theta_0(\boldsymbol{x})|\boldsymbol{x}\big\} = f_{\varepsilon|\boldsymbol{X}}\big(v|\boldsymbol{x}\big)$, but

this property does not hold for the kernel estimators

$$\hat{f}_{\varepsilon|\boldsymbol{X}}(0|\boldsymbol{x}) = n^{-1}\sum_{i=1}^{n} g_{h_0}\left(\hat{\varepsilon}_i\right) L_{\bar{h}}(\boldsymbol{x}-\boldsymbol{X}_i)/\hat{f}_{\boldsymbol{X}}(\boldsymbol{x}), \tag{3.3.10}$$

$$\hat{f}_{Y|\boldsymbol{X}}(\hat{\theta}_n(\boldsymbol{x})|\boldsymbol{x}) = n^{-1}\sum_{i=1}^{n} g_{h_1}\left(Y_i-\hat{\theta}_n(\boldsymbol{x})\right) L_{\tilde{h}}(\boldsymbol{x}-\boldsymbol{X}_i)/\hat{f}_{\boldsymbol{X}}(\boldsymbol{x}), \tag{3.3.11}$$

of the conditional density functions. In general $\hat{f}_{\varepsilon|\boldsymbol{X}}(0|\boldsymbol{x}) \neq \hat{f}_{Y|\boldsymbol{X}}(\hat{\theta}_n(\boldsymbol{x})|\boldsymbol{x})$ in $\boldsymbol{x}$ although both estimates are asymptotically equivalent. In applications the two estimators can differ substantially due to the bandwidth selection because for data-driven bandwidths we usually have $h_0 \neq h_1$. For example, if a common method for bandwidth selection such as a rule-of-thumb is used, h_1 will tend to be larger than h_0 since the sample variance of Y_i tends to be larger than that of $\hat{\varepsilon}_i$. Given that the same kernels are applied, it happens often that $\hat{f}_{Y|\boldsymbol{X}}(\hat{\theta}_n(\boldsymbol{x})|\boldsymbol{x}) > f_{Y|\boldsymbol{X}}(\theta_0(\boldsymbol{x})|\boldsymbol{x})$, even if $\hat{\theta}_n(\boldsymbol{x})$ is usually very close to $\theta_0(\boldsymbol{x})$. To correct such abnormality, we are motivated to set $h_1 = h_0$ which is the rule-of-thumb bandwidth of $\hat{f}_{\varepsilon|\boldsymbol{x}}(v|\boldsymbol{x})$ in (3.3.11). As the result, it leads to a more rough estimate for $\hat{f}_{Y|\boldsymbol{X}}(\hat{\theta}_n(\boldsymbol{x})|\boldsymbol{x})$.

In order to exploit the roughness of $\hat{f}_{Y|\boldsymbol{X}}(\hat{\theta}_n(\boldsymbol{x})|\boldsymbol{x})$ while making the CC as narrow as possible, we develop a trick depending on

$$\frac{\hat{f}_{Y|\boldsymbol{X}}\left\{\hat{\theta}_n(\boldsymbol{x})|\boldsymbol{x}\right\}}{\hat{f}_{\varepsilon|\boldsymbol{X}}(0|\boldsymbol{x})} = \frac{h_0}{h_1}\frac{\sum_{i=1}^{n} g_{h_1}\left(\left\{Y_i-\hat{\theta}_n(\boldsymbol{x})\right\}/h_1\right) L_{\tilde{h}}(\boldsymbol{x}-\boldsymbol{X}_i)}{\sum_{i=1}^{n} g_{h_0}\left(\hat{\varepsilon}_i/h_0\right) L_{\bar{h}}(\boldsymbol{x}-\boldsymbol{X}_i)}. \tag{3.3.12}$$

As $n \to \infty$, (3.3.12) converges to 1. If we impose $h_0 = h_1$, as the multiple h_0/h_1 vanishes, (3.3.12) captures the deviation of the two estimators without the difference of the bandwidth in the way. In particular, the bandwidth $h_0 = h_1$ is selected as the rule-of-thumb bandwidth for $\hat{f}_{\varepsilon|\boldsymbol{X}}(y|\boldsymbol{x})$. This makes $\hat{f}_{\varepsilon|\boldsymbol{X}}(y|\boldsymbol{x})$ larger and thus leads to a narrower CC, as will be more clear below.

We propose the alternative bootstrap confidence corridor for quantile estimator:

$$\left\{\theta : \sup_{\boldsymbol{x}\in\mathcal{D}} \left|\sqrt{\hat{f}_{\boldsymbol{X}}(\boldsymbol{x})}\hat{f}_{Y|\boldsymbol{X}}\left\{\hat{\theta}_n(\boldsymbol{x})|\boldsymbol{x}\right\}\left[\hat{\theta}_n(\boldsymbol{x})-\theta(\boldsymbol{x})\right]\right| \leq \xi_\alpha^\dagger\right\},$$

where $\xi_\alpha^\dagger$ satisfies

$$\mathrm{P}^*\left(\sup_{\boldsymbol{x}\in\mathcal{D}}\left|\hat{f}_{\boldsymbol{X}}(\boldsymbol{x})^{-1/2}\frac{\hat{f}_{Y|\boldsymbol{X}}\left\{\hat{\theta}_n(\boldsymbol{x})|\boldsymbol{x}\right\}}{\hat{f}_{\varepsilon|\boldsymbol{X}}(0|\boldsymbol{x})}\left[A_n^*(\boldsymbol{x})-\mathsf{E}^*A_n^*(\boldsymbol{x})\right]\right| \leq \xi_\alpha^\dagger\right) = 1-\alpha. \tag{3.3.13}$$

Note that the probability on the left-hand side of (3.3.13) can again be approximated by a Gumbel distribution function asymptotically, which follows by Theorem 3.3.1.

3.4 A simulation study

In this section we investigate the methods described in the previous sections by means of a simulation study. We construct confidence corridors for quantiles and

expectiles for different levels τ and use the quartic (product) kernel. The performance of our methods is compared to the performance of the method proposed by Hall and Horowitz (2013) at the end of this section. For the confidence based on asymptotic distribution theory, we use the rule of thumb bandwidth chosen from the `R` package `np`, and then rescale it as described in Yu and Jones (1998), finally multiply it by $n^{-0.05}$ for undersmoothing. The sample sizes are given by $n = 100, 300$ and 500, so the undersmoothing multiples are 0.794, 0.752 and 0.733 respectively. We take 20×20 equally distant grids in $[0.1, 0.9]^2$ and estimate quantile or expectile functions pointwisely on this set of grids. In the quantile regression bootstrap CC, the bandwidth h_1 used for estimating $\hat{f}_{Y|\boldsymbol{X}}(y|\boldsymbol{x})$ is chosen to be the rule-of-thumb bandwidth of $\hat{f}_{\varepsilon|\boldsymbol{X}}(0|\boldsymbol{x})$ and multiplied by a multiple 1.5. This would give slightly wider CCs.

		Homogeneous			**Heterogeneous**		
Method	n	$\tau = 0.5$	$\tau = 0.2$	$\tau = 0.8$	$\tau = 0.5$	$\tau = 0.2$	$\tau = 0.8$
				$\sigma_0 = 0.2$			
	100	.000(0.366)	.109(0.720)	.104(0.718)	.000(0.403)	.120(0.739)	.122(0.744)
	300	.000(0.304)	.130(0.518)	.133(0.519)	.002(0.349)	.136(0.535)	.153(0.537)
	500	.000(0.262)	.117(0.437)	.142(0.437)	.008(0.296)	.156(0.450)	.138(0.450)
				$\sigma_0 = 0.5$			
	100	.070(0.890)	.269(1.155)	.281(1.155)	.078(0.932)	.300(1.193)	.302(1.192)
Asympt.	300	.276(0.735)	.369(0.837)	.361(0.835)	.325(0.782)	.380(0.876)	.394(0.877)
	500	.364(0.636)	.392(0.711)	.412(0.712)	.381(0.669)	.418(0.743)	.417(0.742)
				$\sigma_0 = 0.7$			
	100	.160(1.260)	.381(1.522)	.373(1.519)	.155(1.295)	.364(1.561)	.373(1.566)
	300	.438(1.026)	.450(1.109)	.448(1.110)	.481(1.073)	.457(1.155)	.472(1.152)
	500	.533(0.888)	.470(0.950)	.480(0.949)	.564(0.924)	.490(0.984)	.502(0.986)
				$\sigma_0 = 0.2$			
	100	.325(0.676)	.784(0.954)	.783(0.954)	.409(0.717)	.779(0.983)	.778(0.985)
	300	.442(0.457)	.896(0.609)	.894(0.610)	.580(0.504)	.929(0.650)	.922(0.649)
	500	.743(0.411)	.922(0.502)	.921(0.502)	.839(0.451)	.950(0.535)	.952(0.536)
				$\sigma_0 = 0.5$			
	100	.929(1.341)	.804(1.591)	.818(1.589)	.938(1.387)	.799(1.645)	.773(1.640)
Bootst.	300	.950(0.920)	.918(1.093)	.923(1.091)	.958(0.973)	.919(1.155)	.923(1.153)
	500	.988(0.861)	.968(0.943)	.962(0.942)	.990(0.902)	.962(0.986)	.969(0.987)
				$\sigma_0 = 0.7$			
	100	.976(1.811)	.817(2.112)	.808(2.116)	.981(1.866)	.826(2.178)	.809(2.176)
	300	.986(1.253)	.919(1.478)	.934(1.474)	.983(1.308)	.930(1.537)	.920(1.535)
	500	.996(1.181)	.973(1.280)	.968(1.278)	.997(1.225)	.969(1.325)	.962(1.325)

Table 3.4.1: Nonparametric quantile model coverage probabilities. The nominal coverage is 95%. The number in the parentheses is the volume of the confidence corridor. The asymptotic method corresponds to the asymptotic quantile regression CC and bootstrap method corresponds to quantile regression bootstrap CC.

		Homogeneous			Heterogeneous		
Method	n	$\tau = 0.5$	$\tau = 0.2$	$\tau = 0.8$	$\tau = 0.5$	$\tau = 0.2$	$\tau = 0.8$
				$\sigma_0 = 0.2$			
	100	.000(0.428)	.000(0.333)	.000(0.333)	.000(0.463)	.000(0.362)	.000(0.361)
	300	.049(0.341)	.000(0.273)	.000(0.273)	.079(0.389)	.001(0.316)	.002(0.316)
	500	.168(0.297)	.000(0.243)	.000(0.243)	.238(0.336)	.003(0.278)	.002(0.278)
				$\sigma_0 = 0.5$			
	100	.007(0.953)	.000(0.776)	.000(0.781)	.007(0.997)	.000(0.818)	.000(0.818)
Asympt.	300	.341(0.814)	.019(0.708)	.017(0.709)	.355(0.862)	.017(0.755)	.018(0.754)
	500	.647(0.721)	.067(0.645)	.065(0.647)	.654(0.759)	.061(0.684)	.068(0.684)
				$\sigma_0 = 0.7$			
	100	.012(1.324)	.000(1.107)	.000(1.107)	.010(1.367)	.000(1.145)	.000(1.145)
	300	.445(1.134)	.021(1.013)	.013(1.016)	.445(1.182)	.017(1.062)	.016(1.060)
	500	.730(1.006)	.062(0.928)	.078(0.929)	.728(1.045)	.068(0.966)	.066(0.968)
				$\sigma_0 = 0.2$			
	100	.686(2.191)	.781(2.608)	.787(2.546)	.706(2.513)	.810(2.986)	.801(2.943)
	300	.762(0.584)	.860(0.716)	.876(0.722)	.788(0.654)	.877(0.807)	.887(0.805)
	500	.771(0.430)	.870(0.533)	.875(0.531)	.825(0.516)	.907(0.609)	.904(0.615)
				$\sigma_0 = 0.5$			
	100	.886(5.666)	.906(6.425)	.915(6.722)	.899(5.882)	.927(6.667)	.913(6.571)
Bootst.	300	.956(1.508)	.958(1.847)	.967(1.913)	.965(1.512)	.962(1.866)	.969(1.877)
	500	.968(1.063)	.972(1.322)	.972(1.332)	.972(1.115)	.971(1.397)	.974(1.391)
				$\sigma_0 = 0.7$			
	100	.913(7.629)	.922(8.846)	.935(8.643)	.929(8.039)	.935(9.057)	.932(9.152)
	300	.969(2.095)	.969(2.589)	.971(2.612)	.974(2.061)	.972(2.566)	.979(2.604)
	500	.978(1.525)	.976(1.881)	.967(1.937)	.981(1.654)	.978(1.979)	.974(2.089)

Table 3.4.2: Nonparametric expectile model coverage probability. The nominal coverage is 95%. The number in the parentheses is the volume of the confidence corridor. The asymptotic method corresponds to the asymptotic expectile regression CC and bootstrap method corresponds to expectile regression bootstrap CC.

The data are generated from the normal regression model

$$Y_i = f(X_{1,i}, X_{2,i}) + \sigma(X_{1,i}, X_{2,i})\varepsilon_i, \quad i = 1, \ldots, n \tag{3.4.1}$$

where the independent variables (X_1, X_2) follow a joint uniform distribution taking values on $[0,1]^2$, $\mathrm{Cov}(X_1, X_2) = 0.2876$, $f(X_1, X_2) = \sin(2\pi X_1) + X_2$, and ε_i are independent standard Gaussian random variables. For both quantile and expectile, we look at three quantiles of the distribution, namely $\tau = 0.2, 0.5, 0.8$. The set of grid point is $G \times G$ where G is the set of 20 equidistant grids on univariate interval $[0.1, 0.9]$. Thus, the grid size is $|G \times G| = 400$.

In the homogeneous model, we take $\sigma(X_1, X_2) = \sigma_0$, for $\sigma_0 = 0.2, 0.5, 0.7$. In the heterogeneous model, we take $\sigma(X_1, X_2) = \sigma_0 + 0.8X_1(1 - X_1)X_2(1 - X_2)$. 2000 simulation runs are carried out to estimate the coverage probability.

The upper part of Table 3.4.1 shows the coverage probability of the asymptotic CC for nonparametric quantile regression functions. It can be immediately seen that

the asymptotic CC performs very poorly, especially when n is small. A comparison of the results with those of one-dimensional asymptotic simultaneous confidence bands derived in Claeskens and Van Keilegom (2003) or Fan and Liu (2013), shows that the accuracy in the two-dimensional case is much worse. Much to our surprise, the asymptotic CC performs better in the case of $\tau = 0.2, 0.8$ than in the case of $\tau = 0.5$. On the other hand, it is perhaps not so amazing to see that asymptotic CCs behave similarly under both homogeneous and heterogeneous models. As a final remark about the asymptotic CC we mention that it is highly sensitive with respect to σ_0. Increasing values of σ_0 yields larger CC, and this may lead to greater coverage probability.

The lower part of Table 3.4.1 shows that the bootstrap CCs for nonparametric quantile regression functions yield a remarkable improvement in comparison to the asymptotic CC. For the bootstrap CC, the coverage probabilities are in general close to the nominal coverage of 95%. The bootstrap CCs are usually wider, and getting narrower when n increases. Such phenomenon can also be found in the simulation study of Claeskens and Van Keilegom (2003). Bootstrap CCs are less sensitive than asymptotic CCs with respect to the choice σ_0, which is also considered as an advantage. Finally, we note that the performance of bootstrap CCs does not depend on which variance specification is used too.

The upper part of Table 3.4.2 shows the coverage probabiltiy of the CC for nonparametric expectile regression functions. The results are similar to the case of quantile regression. The asymptotic CCs do *not* give accurate coverage probabilities. For example in some cases like $\tau = 0.2$ and $\sigma_0 = 0.2$, not a single simulation in the 2000 iterations yields a case where surface is completely covered by the asymptotic CC.

The lower part of Table 3.4.2 shows that bootstrap CCs for expectile regression give more accurate approximates to the nominal coverage than the asymptotic CCs. One can see in the parenthesis that the volumes of the bootstrap CCs are significantly larger than those of the asymptotic CCs, especially for small n.

Table 3.4.3 presents the proportion in the 2000 iterations which covers 95% of the 400 grid points, using the bootstrap method proposed in Hall and Horowitz (2013)(abbreviated as HH) for nonparametric mean regression at $d = 2$. HH derived an expansion for the bootstrap bias and established a somewhat different way to construct confidence bands without the use of extreme value theory. It is worth noting that their bands are uniform with respect to a fixed but unspecified portion of $(1 - \xi) \cdot 100\%$ (smaller than 100%) of grid points, while in our approach the uniformity is achieved on the whole set of grids.

The simulation model is (3.4.1) with the same homogeneous and heterogeneous variance specifications as before. We choose three levels of $\xi = 0.005, 0.05$ and 0.1. It is suggested in HH that $\xi = 0.1$ is usually sufficient in univariate nonparametric mean regression $d = 1$. Note that $\xi = 0.005$ corresponds to the second smallest pointwise quantile $\hat{\beta}(\boldsymbol{x}, 0.05)$ in the notation of HH, given that our grid size is 400. This is close to the uniform CC in our sense. The simulation model associated with the Table 3.4.3 is the same with that of the case $\tau = 0.5$ in the bootstrap part of

	Homogeneous			Heterogeneous		
n	$\xi = 0.005$	$\xi = 0.05$	$\xi = 0.1$	$\xi = 0.005$	$\xi = 0.05$	$\xi = 0.1$
			$\sigma_0 = 0.2$			
100	.693(3.027)	.529(1.740)	.319(1.040)	.680(3.452)	.546(2.051)	.332(1.224)
300	.891(0.580)	.748(0.365)	.642(0.323)	.907(0.667)	.798(0.414)	.698(0.364)
500	.886(0.335)	.770(0.265)	.678(0.244)	.896(0.379)	.789(0.298)	.699(0.274)
			$\sigma_0 = 0.5$			
100	.720(7.264)	.611(4.489)	.394(2.686)	.729(7.594)	.616(4.676)	.414(2.829)
300	.945(1.423)	.849(0.859)	.755(0.746)	.940(1.511)	.854(0.912)	.760(0.791)
500	.944(0.795)	.846(0.600)	.750(0.548)	.937(0.833)	.839(0.632)	.751(0.577)
			$\sigma_0 = 0.7$			
100	.730(10.183)	.634(6.411)	.430(3.853)	.752(10.657)	.658(6.577)	.441(3.923)
300	.936(1.995)	.854(1.197)	.751(1.037)	.951(2.091)	.875(1.256)	.772(1.086)
500	.933(1.098)	.854(0.831)	.774(0.758)	.938(1.145)	.853(0.865)	.770(0.789)

Table 3.4.3: Proportion in 2000 iteration that the coverage of $\geq 95\%$ grid points for nonparametric mean model, using the bootstrap method of Hall and Horowitz (2013). The nominal coverage is 95%. The number in the parentheses is the volume of the confidence corridor.

Table 3.4.1 and Table 3.4.2, because in case of the normal distribution the median equals the mean and $\tau = 0.5$ expectile is exactly the mean. However, one should be aware that our coverage probabilities are more stringent because we check the coverage at every point in the set of grids, rather than only 95% of the points (we refer it as *complete coverage*). Hence, the complete coverage probability of HH will be lower than the proportion of 95% coverage shown in Table 3.4.3. The proportion of 95% coverage should therefore be viewed as an upper bound for the complete coverage.

We summarize our findings as follows. Firstly the proportion of 95% coverage in general present similar patterns as shown in Table 3.4.1 and 3.4.2. The coverage improves when n and σ_0 get larger, and the volume of the band decreases as n increases and increases when σ_0 increases. The homogeneous and heterogeneous model yield similar performance. Comparing with the univariate result in HH, it is found that the proportion of coverage tends to perform worse than that in HH under the same sample size. This is due to the curse of dimensionality, the estimation of a bivariate function is less accurate than that of an univariate function. As the result, a more conservative ξ has to be applied. If we compare Table 3.4.3 to the bootstrap part of 3.4.1 with $\tau = 0.5$, it can be seen that our complete coverage probabilities are comparable to the proportion of 95% coverage at the case $\xi = 0.005$, though in the case of $\sigma_0 = 0.2$ our CC does not perform very well. However, the volumes of our CC are much less than that of HH in the cases of small n and moderate and large σ_0. This suggests that our CC is more efficient. Finally, the proportion of 95% coverage at $\xi = 0.005$ in Table 3.4.3 is similar to the complete coverage probability in bootstrap part of 3.4.2 with $\tau = 0.5$, but when sample size is small, the volume

of our CC is smaller.

3.5 Application: a treatment effect study

The classical application of the proposed method consists in testing the hypothetical functional form of the regression function. Nevertheless, the proposed method can also be applied to test for a quantile treatment effect (see Koenker; 2005) or to test for conditional stochastic dominance (CSD) as investigated in Delgado and Escanciano (2013). In this section we shall apply the new method to test these hypotheses for data collected from a real government intervention.

The estimation of the quantile treatment effect (QTE) recovers the heterogeneous impact of intervention on various points of the response distribution. To define QTE, given vector-valued exogenous variables $\boldsymbol{X} \in \boldsymbol{\mathcal{X}}$ where $\boldsymbol{\mathcal{X}} \subset \mathbb{R}^d$, suppose Y_0 and Y_1 are response variables associated with the control group and treatment group, and let $F_{0|\boldsymbol{X}}$ and $F_{1|\boldsymbol{X}}$ be the conditional distribution for Y_0 and Y_1, the QTE at level τ is defined by

$$\Delta_\tau(\boldsymbol{x}) \stackrel{\text{def}}{=} Q_{1|\boldsymbol{X}}(\tau|\boldsymbol{x}) - Q_{0|\boldsymbol{X}}(\tau|\boldsymbol{x}), \quad \boldsymbol{x} \in \boldsymbol{\mathcal{X}}, \tag{3.5.1}$$

where $Q_{0|\boldsymbol{X}}(y|\boldsymbol{x})$ and $Q_{1|\boldsymbol{X}}(y|\boldsymbol{x})$ are the conditional quantile of Y_0 given $\boldsymbol{X}$ and Y_1 given $\boldsymbol{X}$, respectively. This definition corresponds to the idea of horizontal distance between the treatment and control distribution functions appearing in Doksum (1974) and Lehmann (1975).

A related concept in measuring the efficiency of a treatment is the so called "conditional stochastic dominance". Y_1 conditionally stochastically dominates Y_0 if

$$F_{1|\boldsymbol{X}}(y|\boldsymbol{x}) \leq F_{0|\boldsymbol{X}}(y|\boldsymbol{x}) \quad \text{a.s. for all } (y, \boldsymbol{x}) \in (\mathcal{Y}, \boldsymbol{\mathcal{X}}), \tag{3.5.2}$$

where $\mathcal{Y}$, $\boldsymbol{\mathcal{X}}$ are domains of Y and $\boldsymbol{X}$. For example, if Y_0 and Y_1 stand for the income of two groups of people G_0 and G_1, (3.5.2) means that the distribution of Y_1 lies on the right of that of Y_0, which is equivalent to saying that at a given $0 < \tau < 1$, the τ-quantile of Y_1 is greater than that of Y_0. Hence, we could replace the testing problem (3.5.2) by

$$Q_{1|\boldsymbol{X}}(\tau|\boldsymbol{x}) \geq Q_{0|\boldsymbol{X}}(\tau|\boldsymbol{x}) \quad \text{for all } 0 < \tau < 1 \text{ and } \boldsymbol{x} \in \boldsymbol{\mathcal{X}}. \tag{3.5.3}$$

Comparing (3.5.3) and (3.5.1), one would find that (3.5.3) is just a uniform version of the test $\Delta_\tau(\boldsymbol{x}) \geq 0$ over $0 < \tau < 1$.

The method that we introduced in this paper is suitable for testing a hypothesis like $\Delta_\tau(\boldsymbol{x}) = 0$ where $\Delta_\tau(\boldsymbol{x})$ is defined in (3.5.1). One can construct CCs for $Q_{1|\boldsymbol{X}}(\tau|\boldsymbol{x})$ and $Q_{0|\boldsymbol{X}}(\tau|\boldsymbol{x})$ respectively, and then check if there is overlap between the two confidence regions. One can also extend this idea to test (3.5.3) by building CCs for several selected levels τ.

We use our method to test the effectiveness of the National Supported Work (NSW) demonstration program, which was a randomized, temporary employment

program initiated in 1975 with the goal to provide work experience for individuals who face economic and social problems prior to entering the program. The data have been widely applied to examine techniques which estimate the treatment effect in a nonexperimental setting. In a pioneer study, LaLonde (1986) compares the treatment effect estimated from the experimental NSW data with that implied by nonexperimental techniques. Dehejia and Wahba (1999) analyse a subset of Lalonde's data and propose a new estimation procedure for nonexperimental treatment effect giving more accurate estimates than Lalonde's estimates. The paper that is most related to our study is Delgado and Escanciano (2013). These authors propose a test for hypothesis (3.5.2) and apply it to Lalonde's data, in which they choose "age" as the only conditional covariate and the response variable being the increment of earnings from 1975 to 1978. They cannot reject the null hypothesis of nonnegative treatment effect on the earnings growth.

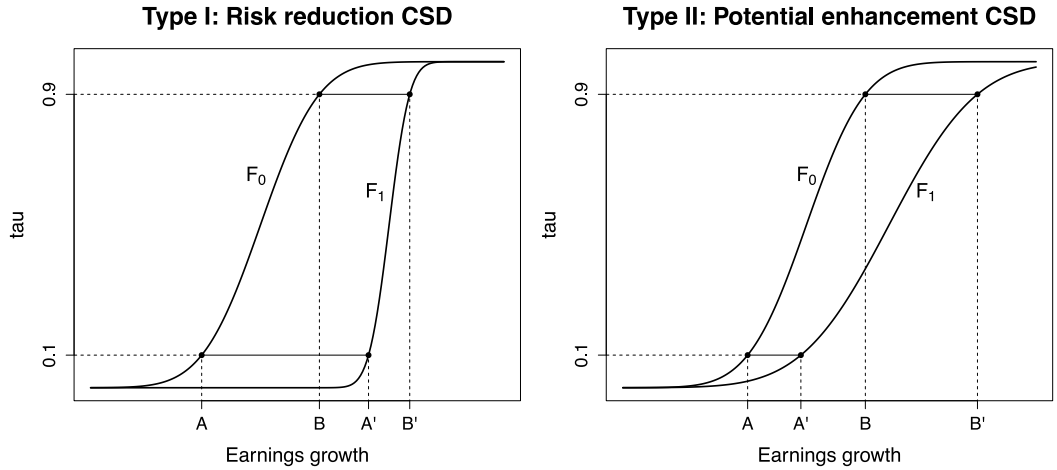


Figure 3.5.1: The illustrations for the two possible types of stochastic dominance. In the left figure, the 0.1 quantile improves (downside risk reduction) more dramatically than the 0.9 quantile (upside potential increase), as the distance between A and A' is greater than that between B and B'. For the right picture the interpretation is just the opposite.

The previous literature, however, has not addressed an important question. We shall depict this question by two pictures. In Figure 3.5.1, it is obvious that Y_1 stochastically dominates Y_0 in both pictures, but significant differences can be seen between the two scenarios. For the left one, the 0.1 quantile improves more dramatically than the 0.9 quantile, as the distance between A and A' is greater than that between B and B'. In usual words, the gain of the 90% lower bound of the earnings growth is more than that of the 90% upper bound of the earnings growth after the treatment. "90% lower bound of the earnings growth" means the probability that the earnings growth is above the bound is 90%. This suggests that the treatment induces greater reduction in downside risk but less increase in the upside potential in the earnings growth. For the right picture the interpretation is just the opposite.

To see which type of stochastic dominance the NSW demonstration program belongs to, we apply the same data as Delgado and Escanciano (2013) for testing the hypothesis of positive quantile treatment effect for several quantile levels τ. The data consist of 297 treatment group observations and 423 control group observations. The response variable Y_0 (Y_1) denotes the difference in earnings of control (treatment) group between 1978 (year of postintervention) and 1975 (year of preintervention). We first apply common statistical procedures to describe the distribution of these two variables. Figure 3.5.2 shows the unconditional densities and distribution function. The cross-validated bandwidth for $\hat{f}_0(y)$ is 2.273 and 2.935 for $\hat{f}_1(y)$. The left figure of Figure 3.5.2 shows the unconditional densities of the income difference for treatment group and control group. The density of the treatment group has heavier tails while the density of the control group is more concentrated around zero. The right figure shows that the two unconditional distribution functions are very close on the left of the 50% percentile, and slight deviation appears when the two distributions are getting closer to 1. Table 3.5.1 shows that, though the differences are small, but the quantiles of the unconditional cdf of treatment group are mildly greater than that of the control group for each chosen τ. The two-sample Kolmogorov-Smirnov and Cramér-von Mises tests, however, yield results shown in the Table 3.5.2 which cannot reject the null hypothesis that the empirical cdfs for the two groups are the same with confidence levels 1% or 5%.

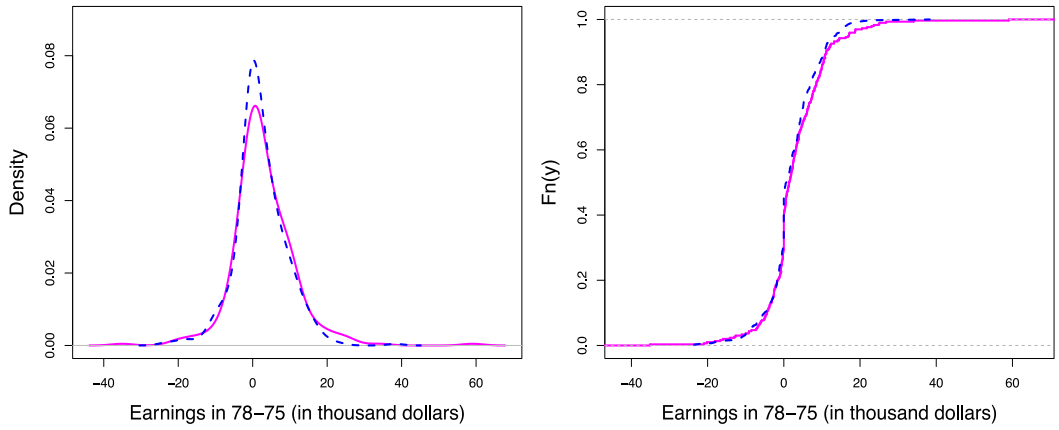


Figure 3.5.2: Unconditional empirical density function (left) and distribution function (right) of the difference of earnings from 1975 to 1978. The dashed line is associated with the control group and the solid line is associated with the treatment group.

Next we apply our test on quantile regression to evaluate the treatment effect. In order to compare with Delgado and Escanciano (2013), we first focus on the case of a one-dimensional covariate. The first covariate X_{1i} is the age. The second covariate X_{2i} is the number of years of schooling. The sample values of schooling years lie in the range of $[3, 16]$ and age lies between $[17, 55]$. In order to avoid

$\tau(\%)$	10	20	30	50	70	80	90
Treatment	-4.38	-1.55	0.00	1.40	5.48	8.50	11.15
Control	-4.91	-1.73	-0.17	0.74	4.44	7.16	10.56

Table 3.5.1: The unconditional sample quantiles of treatment and control groups.

Type of test	Statistics	p-value
Kolmogorov-Smirnov	0.0686	0.3835
Cramér-von Mises	0.2236	0.7739

Table 3.5.2: The two sample empirical cdf tests results for treatment and control groups.

boundary effect and sparsity of the samples, we look at the ranges [7,13] for schooling years and [19,31] for age. We apply the bootstrap CC method for quantiles $\tau = 0.1, 0.2, 0.3, 0.5, 0.7, 0.8$ and 0.9. We apply the quartic kernel. The cross-validated bandwidths are chosen in the same way as for conditional densities with the `R` package `np`. The resulting bandwidths are (2.2691,2.5016) for the treatment group and (2.7204, 5.9408) for the control group. In particular, for smoothing the data of the treatment group, for $\tau = 0.1$ and 0.9, we enlarge the cross-validated bandwidths by a constant of 1.7; for $\tau = 0.2, 0.3, 0.7, 0.8$, the cross-validated bandwidths are enlarged by constant factor 1.3. These inflated bandwidths are used to handle violent roughness in extreme quantile levels. The bootstrap CCs are computed with 10,000 repetitions. The level of the test is $\alpha = 5\%$.

The results of the two quantile regressions with one-dimensional covariate, and their CCs for various quantile levels are presented in Figure 3.5.3 and 3.5.4. We observe that for all chosen quantile levels the quantile estimates associated to the treatment group lie above that of the control group when age is over certain levels, and particularly for $\tau = 10\%, 50\%, 80\%$ and 90%, the quantile estimates for treatment group exceeds the upper CCs for the quantile estimates of the control group. On the other hand, at $\tau = 10\%$, the quantile estimates for the control group drop below the CC for treatment group for age greater than 27. Hence, the results here show a tendency that both the downside risk reduction and the upside potential enhancement of earnings growth are achieved, as the older individuals benefit the most from the treatment. Note that we observe a heterogeneous treatment effect in age and the weak dominance of the conditional quantiles of the treatment group with respect to those of the control group, i.e., (3.5.3) holds for the chosen quantile levels, which are in line with the findings of Delgado and Escanciano (2013).

We now turn to Figure 3.5.4, where the covariate is the years of schooling. The treatment effect is not significant for conditional quantiles at levels $\tau = 10\%, 20\%$ and 30%. This suggests that the treatment does little to reduce the downside risk of the earnings growth for individuals with various degrees of education. Nonetheless,

we constantly observe that the regression curves of the treatment group rise above that of the control group after a certain level of the years of schooling for quantile levels $\tau = 50\%, 70\%, 80\%$ and 90%. Notice that for $\tau = 50\%$ and 80% the regression curves associated to the treatment group reach the upper boundary of the CC of the control group. This suggests that the treatment effect tends to raise the upside potential of the earnings growth, in particular for those individuals who spent more years in the school. It is worth noting that we also see a heterogeneous treatment effect in schooling years, although the heterogeneity in education is less strong than the heterogeneity in age.

The previous regression analyses separately conditioning on covariates age and schooling years only give a limited view on the performance of the program, we now proceed to the analysis conditioning on the two covariates jointly (X_{1i}, X_{2i}). The estimation settings are similar to the case of univariate covariate. Figure 3.5.5 shows the quantile regression CCs. From a first glance of the pictures, the τ-quantile CC of the treatment group and that of the control group overlap extensively for all τ. We could not find sufficient evidence to reject the null hypothesis that the conditional distribution of treatment group and control group are equivalent.

The second observation obtained from comparing subfigures in Figure 3.5.6, we find that the treatment has larger impact in raising the upper bound of the earnings growth than improving the lower bound. For lower quantile levels $\tau = 10\%, 20\%$ and 30% the solid surfaces uniformly lie inside the CC of the control group, while for $\tau = 50\%, 70\%, 80\%$ and 90%, we see several positive exceedances over the upper boundary of the CC of the control group. Hence, the program tends to do better at raising the upper bound of the earnings growth but does worse at improving the lower bound of the earnings growth. In other words, the program tends to increase the potential for high earnings growth but does little in reducing the risk of negative earnings growth.

Our last conclusion comes from inspecting the shape of the surfaces: conditioning on different levels of years of schooling (age), the treatment effect is heterogeneous in age (years of schooling). The most interesting cases occur when conditioning on high age and high years of schooling. Indeed, when considering the cases of $\tau = 80\%$ and 90%, when conditioning on the years of schooling at 12 (corresponding to finishing the high school), the earnings increment of the treatment group rises above the upper boundary of the CC of the control group. This suggests that the individuals who are older and have more years of schooling tend to benefit more from the treatment.

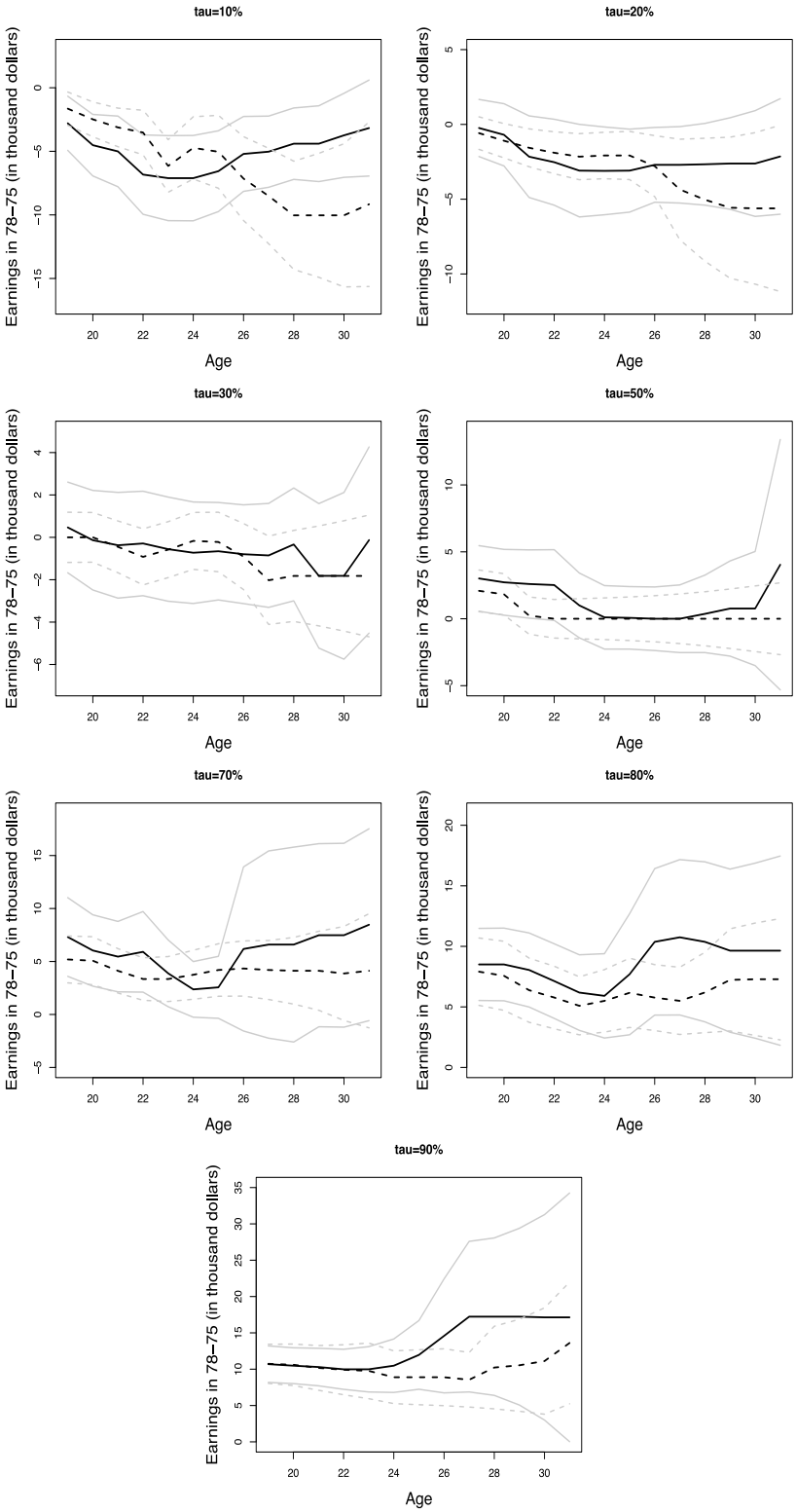


Figure 3.5.3: Nonparametric quantile regression estimates and CCs for the changes in earnings between 1975-1978 as a function of age. The solid dark lines correspond to the conditional quantile of the treatment group and the solid light lines sandwich its CC, and the dashed dark lines correspond to the conditional quantiles of the control group and the solid light lines sandwich its CC.

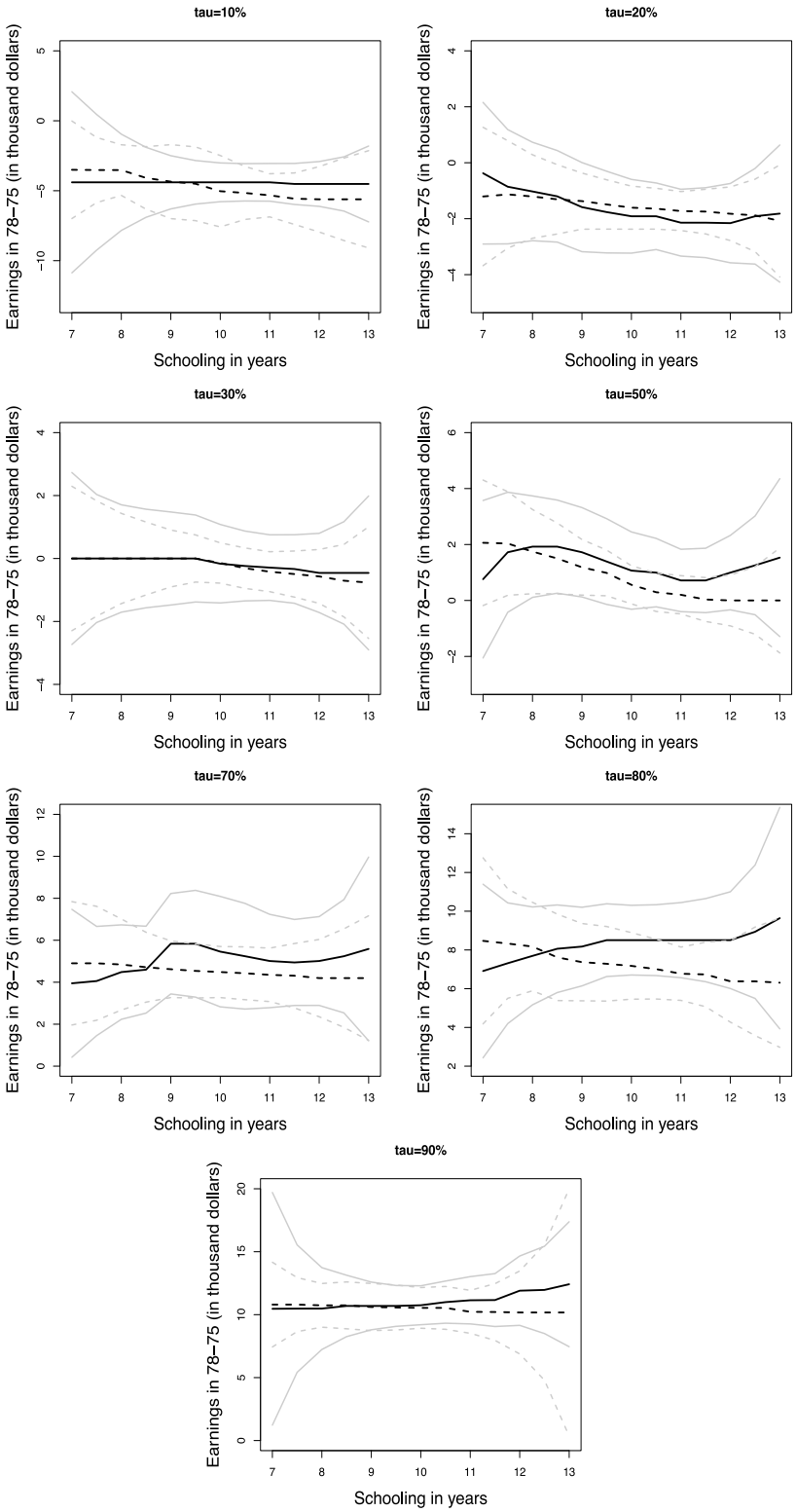
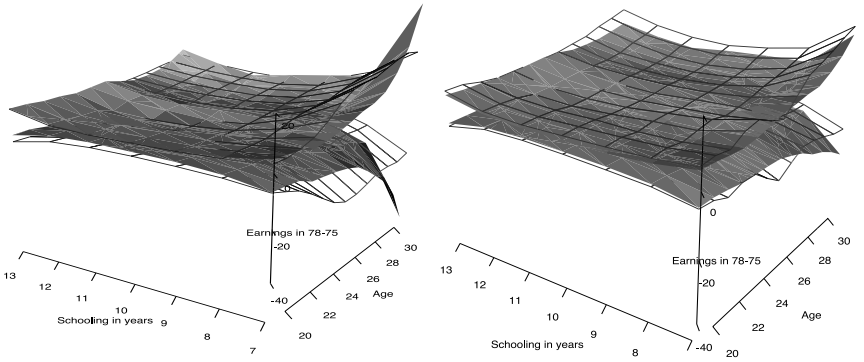
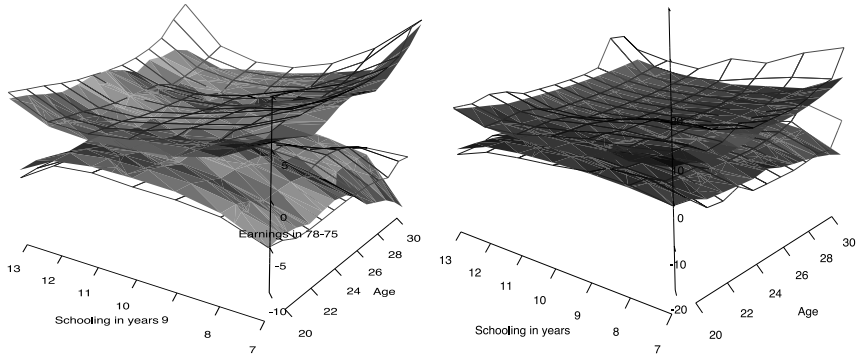


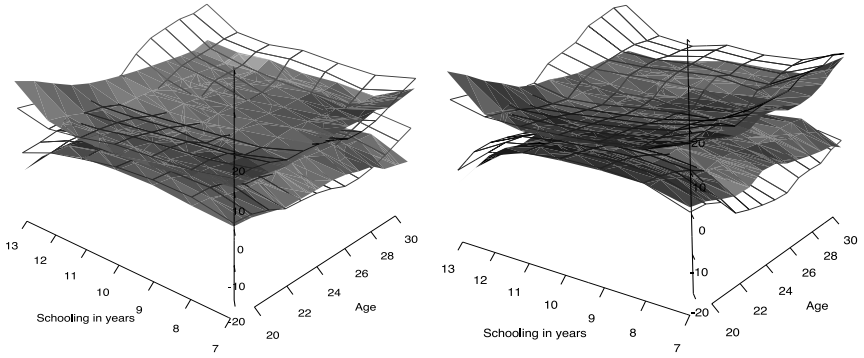
Figure 3.5.4: Nonparametric quantile regression estimates and CCs for the changes in earnings between 1975-1978 as a function of years of schooling. The solid dark lines correspond to the conditional quantile of the treatment group and the solid light lines sandwich its CC, and the dashed dark lines correspond to the conditional quantiles of the control group and the solid light lines sandwich its CC.



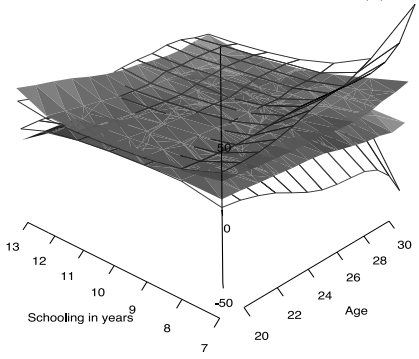
(a) $\tau = 10\%$ (b) $\tau = 20\%$



(c) $\tau = 30\%$ (d) $\tau = 50\%$

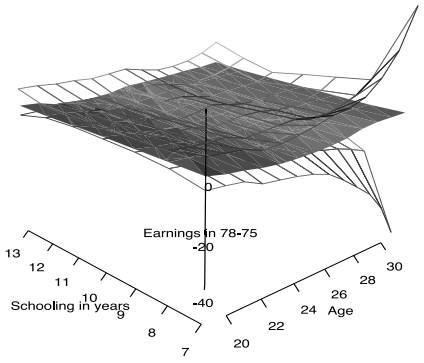


(e) $\tau = 70\%$ (f) $\tau = 80\%$

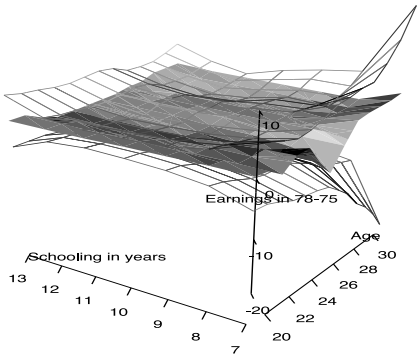


(g) $\tau = 90\%$

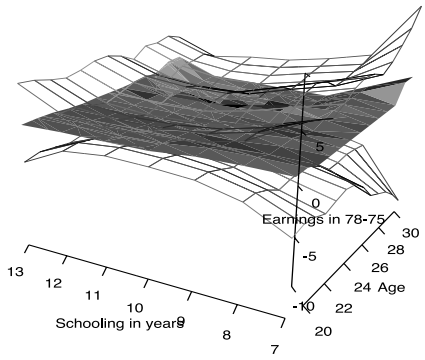
Figure 3.5.5: The CCs for the treatment group and the control group. The net surface corresponds to the control group quantile CC and the solid surface corresponds to the treatment group quantile CC.



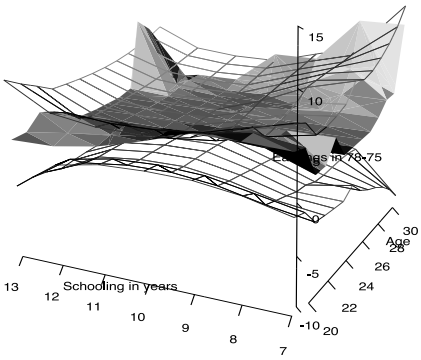
(a) $\tau = 10\%$



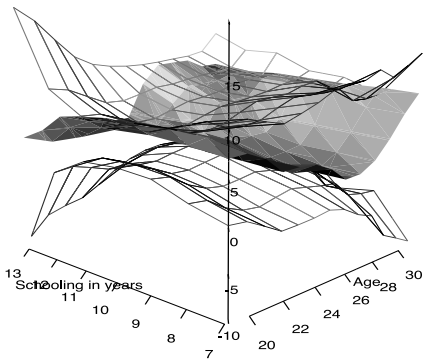
(b) $\tau = 20\%$



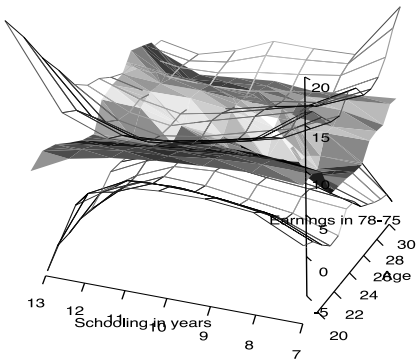
(c) $\tau = 30\%$



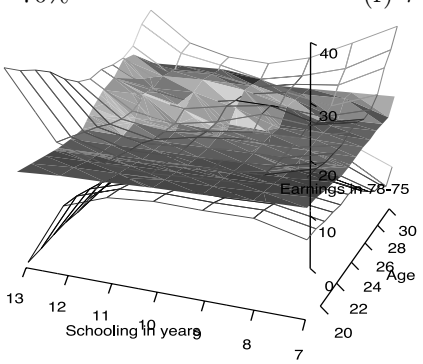
(d) $\tau = 50\%$



(e) $\tau = 70\%$



(f) $\tau = 80\%$



(g) $\tau = 90\%$

Figure 3.5.6: The conditional quantiles (solid surfaces) for the treatment group and the CCs (net surfaces) for the control group.

Chapter 4

FASTEC: Factorisable Sparse Tail Event Curves

4.1 Introduction

High-dimensional multivariate quantile analysis is crucial for many applications, such as risk management and weather analysis. In these applications, quantile functions $q_Y(\tau)$ of random variable Y such that $\mathrm{P}\{Y \leq q_Y(\tau)\} = \tau$ at the "tail" of the distribution, namely at τ close 0 or 1, such as $\tau = 1\%, 5\%$ or $\tau = 95\%, 99\%$, is of great interest. This is because the quantile at level τ can be interpreted as the lower (upper) bound with confidence level $1 - \tau$ (τ) of the possible outcome of a random variable, which can assist the process of decision making for treatment or risk management. Some practical examples:

- Financial risk management: quantiles $q_Y(\tau)$ of asset return with small τ indicates the lower bound of the potential loss, which is of interest of both risk manager and market regulator. In particular, the quantile of asset return with $\tau = 1\%$ is called "value-at-risk". At the same time, this is a high-dimensional problem as there are often several hundreds or thousands of asset returns to be considered.

- Temperature analysis: quantiles at high and small τ give the range of possible temperature variation, which is useful for crop growth or studying climate change. There may be hundreds of weather stations depending on the size of the region being considered.

A global analysis in the behavior of dispersion of high-dimensional random variables can be done based on the observation that the difference of the quantile pair $(q(\tau), q(1-\tau))$ gives a flavor of range, which we refer as *τ-range*. For example $\tau = 25\%$ gives the interquartile range, which is known to be a robust measure of distribution dispersion. The terminology *global* refers to the analysis of the pattern of dispersion of variables, which should be distinguished from the *localized* analysis specialized at a quantile level. While the factors for each of the two quantile allows for modeling asymmetry of distribution, we can detect asymmetric change of

the range of the variables, such as expanding, shrinking, shifting, or shifting while expanding/shrinking, by the sign of loadings and the trend of the factors.

Most previous data analysis method for high-dimensional data emphasizes on the variance and covariance structure of the high-dimensional data, and methods based on that such as principal component analysis can describe the linear dependence in variables when the data are symmetric, in similar scale and no outliers. However, knowing the linear dependence of the random variables does not lead to the knowledge in their lower and/or upper bounds. Moreover, for non-Gaussian and highly asymmetric (skewed) data, the methods based on covariance structure can be highly corrupted if no correction is made.

To see that the information from the covariance and quantiles are not much related, we analayse data simulated from an asymmetric model. The data are simulated with

$$\begin{aligned} Y_{ij} &= \Phi^{-1}(U_{ij})\boldsymbol{X}_i^\top \boldsymbol{\Gamma}_{1,*j}\mathbf{1}(U_{ij} < 0.5), \quad j = 1, ..., 100, \\ Y_{ij} &= \Phi^{-1}(U_{ij})\boldsymbol{X}_i^\top \boldsymbol{\Gamma}_{2,*j}\mathbf{1}(U_{ij} \geq 0.5), \quad j = 101, ..., 200, \end{aligned} \tag{4.1.1}$$

for $i = 1, ..., 500$, where $\{\boldsymbol{X}_i\}$ are i.i.d. from a joint uniform $[0, 1]$ distribution with $\boldsymbol{X}_i \in \mathbb{R}^{200}$, $\{U_{ij}\}$ are i.i.d. uniform $[0, 1]$ over both i and j. $\boldsymbol{\Gamma}_{1,*j}$ and $\boldsymbol{\Gamma}_{2,*j}$ are j column vector of matrices $\boldsymbol{\Gamma}_1, \boldsymbol{\Gamma}_2 \in \mathbb{R}^{p\times m}$, which are of rank 2 and $p = m = 200$. $\Phi(\cdot)$ is the cdf of standard Gaussian distribution. Conditioning on $\boldsymbol{X}_i$, Y_{ij} are independent over j. Notice that the distribution of Y_{ij} are highly asymmetric and skewed, since the first 100 variables are essentially negative and the last 100 are nonnegative. Moreover, the distribution of Y_{ij} is not continuous, since there is nonzero density mass (1/2) at 0.

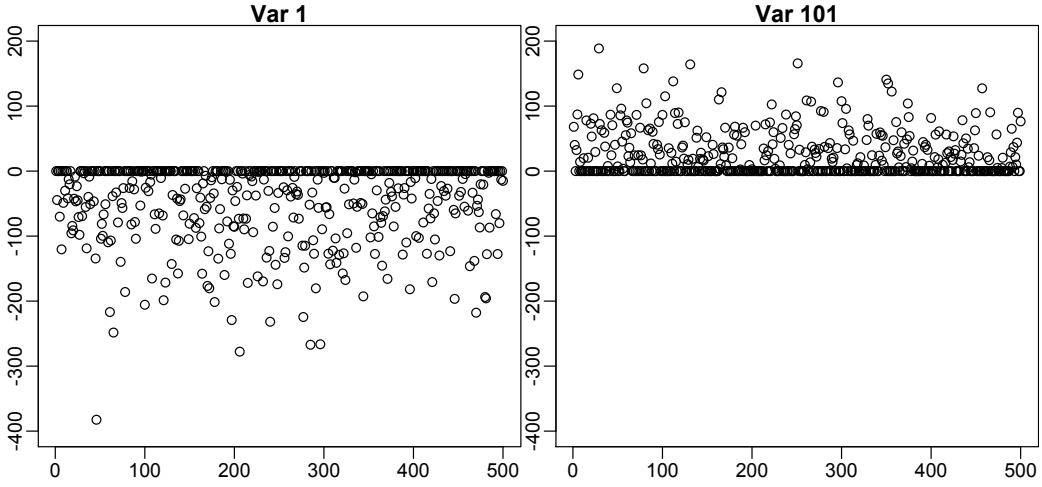


Figure 4.1.1: The variable simulated by (4.1.1). The left is Y_1 bounded above by 0 and the left is Y_{101} bounded below by 0.

The left figure of Figure 4.1.2 is the biplot of PCA on the matrix $\mathbf{Y} = (Y_{ij})$, which suggests that Y_{42} and Y_1 are different variables, and Y_{42} seems to be negatively associated with Y_1 and is perpendicular to Y_{42}. However, the quantile based

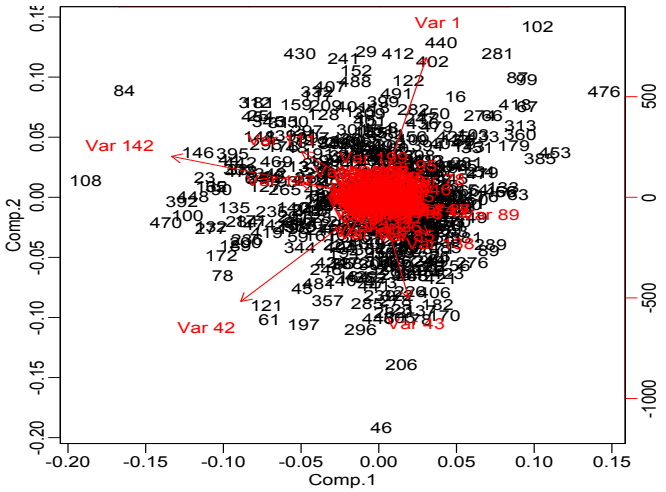


Figure 4.1.2: The PCA biplot on data $\mathbf{Y}$. PCA is based on the covariance and does not capture the pattern in the quantiles of the distribution.

factor analysis (our method) classifies the data with respect to the behavior of their quantiles at the tail ($\tau = 1\%, 99\%$) of the distribution. As the first 100 random variables are similar in their tail behavior (bounded by 0 above), they all lie horizontally close to the x-axis, while the last 100 variables are lying vertically close to the y-axis. The reason for such phenomenon is that PCA takes a *centralized* view and looks at the covariance $\text{Cov}(Y_{ij}, Y_{ik})$ for $j \neq k$, and based on (4.1.1), the inner product of vectors $\mathbf{\Gamma}_{*j}$ and $\mathbf{\Gamma}_{*k}$ plays a big role in it.

Our method, however, looks at the dispersion of the data Y_{ij} from an *uncentralized* view. From the factors and factor loadings in both figures of Figure 4.1.3, the pattern of change in quantiles at 1% and 99% and in τ-range can be determined. Furthermore, in a classification perspective, the variables close with each other on the right of Figure 4.1.3 have similar pattern in the change of the τ-range.

In this paper, we estimate the conditional quantile for high-dimensional data with covariates which is factorisable. This method allows for the global analysis of τ-range or localized analysis of a specific quantile of high-dimensional data, and is more robust to outliers and is capable of capturing the asymmetric distributional dispersion in the data. The key intermediate step of implementation is to estimate conditional quantiles for multivariate responses, which is done via the nuclear norm regularized *multivariate quantile regression*(MQR), in which the we *factorise* the covariates and then using the factors to interpret the data. To handle high-dimensional data, we assume that the coefficient matrix is of low rank. The detail is discussed in later sections.

The low-rank regression has been applied to handle overparametrization and sparse sample size. Reduced-rank multivariate regression is of interest in a wide variety of science fields for cross-sectional data. The earliest work dates back to Anderson (1951) in which the relation between a set of macroeconomic variables and set of manipulable noneconomic variables was considered. Izenman (1975) for-

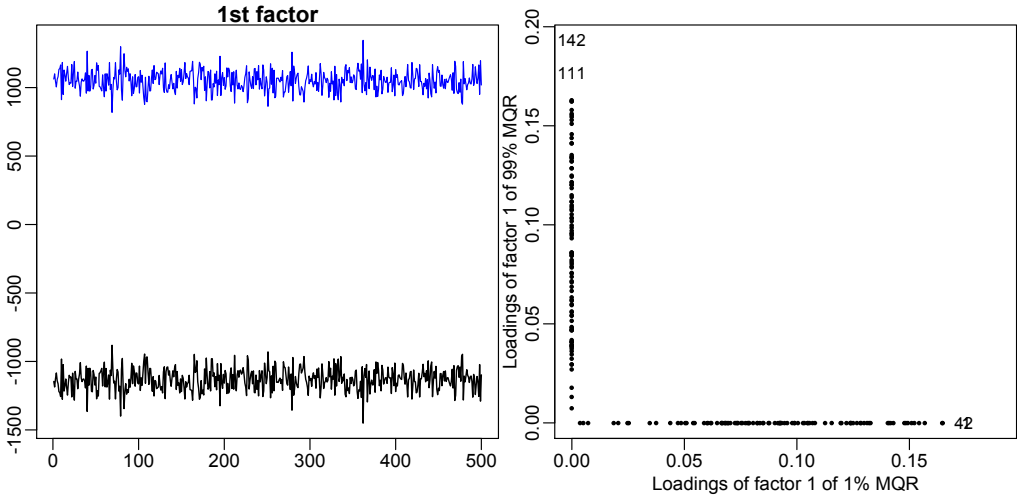


Figure 4.1.3: The first factor of 1% (black) and 99% (blue) quantiles of data $\mathbf{Y}$(left) and the factor loadings(right). Variables have close distance on the right figure have similar change in τ-range, $\tau = 1\%$.

mally introduced the term "reduced-rank regression" and anlaysed the model in detail. For more historical accounts, see Reinsel and Velu (1998) among others. The multivariate regression problem focuses on the expected values of the conditional distributions of m response variables, given p-dimensional covariates. The reduced-rank multivariate regression factorizes the covariates into a parsimonious group of r factors, which decompose the variation of the conditional expectations of the response variables and improve the interpretability of the cross-sectional data.

The estimation of the conditional quantiles with low rank covariate matrix involves minimization of the empirical loss based on the "check function" of Koenker and Bassett (1978), with an additional regularization term of nuclear norm. Our model is equivalent to a multi-task quantile regression with low-rank structure. Fan et al. (2013) also consider multi-task quantile regression under transnormal model.

Our contributions are summarized as follows:

1. The factor model for the quantiles of cross-sectional data is proposed;

2. A method of estimation is designed for the nuclear norm regularized non-smooth empirical loss function and its efficiency is $\mathcal{O}(1/\epsilon)$ where ϵ is a given accuracy level;

3. The nonasymptotic risk bounds for the multivariate quantile regression are derived and are illustrated by numerical analyses;

4. A CAViaR modification for financial risk management is demonstrated.

5. A nonparametric curve model is considered for quantile curves and applied on temperature data.

The modification of the Conditional Autoregressive Value-at-Risk (CAViaR) model of Engle and Manganelli (2004) leads to a Sparse Asymmetric Multivariate Conditional Value-at-Risk (SAMCVaR) model. It can be viewed as a multiple factor version of White et al. (2010), but there is no need to identify the factors nor specifying the number of the factors. We apply SAMCVaR to a dataset consisting of banks, insurance companies and financial service firms from around the world between mid 2007 to mid 2010, including the period of financial crisis. Our first finding is the *negative* leverage effect, in the sense that loss leads to the drop of lower quantile factor rather than the rise of upper quantile factor, which is a step further of the classical result that only suggests the loss leading to higher dispersion of the distribution. Moreover, we show the main risk drivers and risk sensitive firms in the crisis period after the beginning of year 2009. Nonparametric quantile curve model is an extension for the linear multivariate quantile regression model. Using the temperature data, we show that the quantile curve model discriminates the two extreme temperature type in Chinese very well.

4.1.1 Related work

Multivariate quantile regression is studied under several different frameworks by previous authors, but none of them considered high-dimensional case. Serfling (2002) gives a survey of this research direction. Suppose $(\boldsymbol{X}_1, \boldsymbol{Y}_1), ..., (\boldsymbol{X}_n, \boldsymbol{Y}_n)$ are i.i.d. copies of $(\boldsymbol{X}, \boldsymbol{Y})$ in $\mathbb{R}^{p+m}$. Koenker and Portnoy (1990) suggested M-estimation in multiresponse linear regression model with weighting matrix. The estimator has an efficient covariance structure, but the estimator fails to be affine equivariant. Chaudhuri (1996), Koltchinskii (1997), and Chakraborty (2003) consider the geometric quantile, which is the minimizer

$$\arg \min_{\mathbf{S} \in \mathbb{R}^{p \times m}} \left\{ \sum_{i=1}^{n} \|\boldsymbol{Y}_i - \mathbf{S}^\top \boldsymbol{X}_i\| + \mathbf{u}^\top (\boldsymbol{Y}_i - \mathbf{S}^\top \boldsymbol{X}_i) \right\}, \tag{4.1.2}$$

where $\mathbf{u} \in B^{m-1} = \{\mathbf{v} \in \mathbb{R}^m : \|\mathbf{v}\| < 1\}$ controls the direction of deviation from the center of the data cloud and $\|\mathbf{u}\|$ measures the magnitude of the deviation; particularly, $\|\mathbf{u}\| = 0$ corresponds to the median of the data cloud and $\|\mathbf{u}\|$ close to 1 corresponds to the tail of the distribution. Another line of literature tries to link quantile regression and data depth of Tukey (1975). Kong and Mizera (2012) estimate quantile halfspace by projecting data on an oriented straight line with unit vector $\mathbf{u}$, and then finding the quantile hyperplane which is perpendicular to the vector $\mathbf{u}$ and coincides with the line at the quantile of the projected data. The quantile halfspace is the space lying above the hyperplane. They show that their quantile halfspace correspond to Tukey's halfspace depth at each chosen unit vector $\mathbf{u}$. However, in practice their method cannot be used to construct the halfspace depth, because that would require estimating uncountably many quantile spaces. Hallin et al. (2010) propose a novel estimation method quantile halfspaces, and show that the upper envelop of the resulting upper quantile halfspaces coincides

with Tukey's halfspace depth and is computable. Asymptotic properties including Bahadur representation are also established in this paper.

High-dimensional multivariate regression (MR) has been extensively studied in recent years, though the non high-dimensional MR has been around for decades. We review some key ingredients of this model. Suppose

$$\boldsymbol{Y}_i = \boldsymbol{\Gamma}^\top \boldsymbol{X}_i + \boldsymbol{\varepsilon}_i, , \tag{4.1.3}$$

where the entries of $\boldsymbol{\varepsilon}_i$ are independent and with mean 0. In order to recover the matrix $\boldsymbol{\Gamma}$, assuming that $\boldsymbol{\varepsilon}_i \sim N(0, \boldsymbol{\Sigma}_{\boldsymbol{\varepsilon}})$, one minimizes the loss (or negative log likelihood) $\operatorname{tr}\left[(\mathbf{Y} - \mathbf{XS})\boldsymbol{\Omega}(\boldsymbol{Y} - \boldsymbol{X}\mathbf{S})^\top\right]$ with respect to matrix $\mathbf{S}$, where $\boldsymbol{\Omega}$ is a weighting matrix. Common choices are $\boldsymbol{\Omega} = \boldsymbol{\Sigma}_{\boldsymbol{\varepsilon}}^{-1}$ or $\mathbf{I}_m$, while the former choice generates the efficient estimator and the later choice only guarantees consistency. An issue of this approach is that it neglects the dependency in the response variables in covariates $\boldsymbol{X}$ (heteroskedasticity). Another issue is overparametrization, since p and m can be large relative to n and one cannot hope to consistently estimate the model. To deal with these two issues, Izenman (1975) proposed the reduced rank approach. For a predetermined integer $r > 0$,

$$\arg \min_{\mathbf{S} \in \mathbb{R}^{p \times m}} \operatorname{tr}\left[(\mathbf{Y} - \mathbf{XS})\boldsymbol{\Omega}(\mathbf{Y} - \mathbf{XS})^\top\right] \quad \text{s.t. } \operatorname{rank}(\mathbf{S}) \leq r.$$

The number of variables unknown is thus reduced to $r \ll \max\{p, m\}$. Reinsel and Velu (1998) gave an explicit review of this approach.

In the traditional approach described above, r has to be determined ex-ante. In more recent developments, Yuan et al. (2007) proposed a penalization approach, in which they estimate the $\boldsymbol{\Gamma}$ matrix by minimizing:

$$\|\mathbf{Y} - \mathbf{X}\boldsymbol{\Gamma}\|_{\mathrm{F}} + \lambda \|\boldsymbol{\Gamma}\|_*, \tag{4.1.4}$$

where $\lambda > 0$ is a constant. They pointed out the connection between the reduced rank model and factor analysis and proved that an estimator $\widehat{\boldsymbol{\Gamma}}$ can be obtained by soft-thresholding the OLS estimator. Bunea et al. (2011) estimate $\boldsymbol{\Gamma}$ by minimizing $\|\mathbf{Y} - \mathbf{X}\boldsymbol{\Gamma}\|_{\mathrm{F}} + \lambda \operatorname{rank}(\boldsymbol{\Gamma})$, and they show nonasymptotic risk bounds for both their estimator and the estimator from minimizing (4.1.4). They also show that both estimators recover the rank of $\boldsymbol{\Gamma}$ with high probability. In high-dimensional setting, Negahban and Wainwright (2011) consider the cases in which $\boldsymbol{\Gamma}$ is either exact low rank or near low rank. For both cases, they obtain nonasymptotic risk bounds for estimating the true $\boldsymbol{\Gamma}$ with nuclear norm penalized estimator $\widehat{\boldsymbol{\Gamma}}$. Negahban et al. (2012) present a unified framework for analyzing high-dimensional M-estimator with differentiable convex loss functions and decomposable penalizing term. Although the nuclear norm is decomposable, the asymmetric absolute loss function for estimating conditional quantiles is not differentiable and cannot be minorized with a quadratic function, so that the framework of Negahban et al. (2012) cannot be directly applied to our problem.

For high-dimensional multi-task quantile regression, Fan et al. (2013) consider the problem under a transnormal model. They estimate transformations of independent variables which simultaneously explain the quantile of each response variable

and make the joint distribution of transformed covariates and response Gaussian. Comparing to their work, our method assumes low-rank structure, but we do not impose any distribution assumption.

4.1.2 Notations of this chapter

The following notations are adopted throughout this paper. Given two scalars x and y, $x \wedge y \stackrel{\text{def}}{=} \min\{x, y\}$ and $x \vee y \stackrel{\text{def}}{=} \max\{x, y\}$. $\mathbf{1}(x \leq 0)$ is an index function, which is equal to 1 when $x \leq 0$ and 0 when $x > 0$. For a vector $\mathbf{v} = (v_1, ..., v_p) \in \mathbb{R}^p$, let $\|\mathbf{v}\|_2 = (\sum_{j=1}^p v_j^2)^{1/2}$ and $\|\mathbf{v}\|_\infty = \max_{j \leq p} |v_j|$ be the vector ℓ_2 and infinity norm. For a matrix $\mathbf{A} = (A_{ij}) \in \mathbb{R}^{p \times m}$, given the singular values of $\mathbf{A}$: $\sigma_1(\mathbf{A}) \geq \sigma_2(\mathbf{A}) \geq ... \geq \sigma_{p \wedge m}(\mathbf{A})$, let $\|\mathbf{A}\| = \max_{1 \leq j \leq \min\{p,m\}} \sigma_j(\mathbf{A})$, $\|\mathbf{A}\|_* = \sum_{j=1}^{\min\{p,m\}} \sigma_j(\mathbf{A})$ and $\|\mathbf{A}\|_{\mathrm{F}} = \sqrt{\sum_{j=1}^{\min\{p,m\}} \sigma_j(\mathbf{A})^2} = \mathrm{tr}(\mathbf{A}\mathbf{A}^\top)^{1/2} = \mathrm{tr}(\mathbf{A}^\top\mathbf{A})^{1/2} = (\sum_{j=1}^p \sum_{k=1}^m A_{ij}^2)^{1/2}$ and be the matrix spectral norm, nuclear norm (or trace norm), Frobenius norm. The jth column vector of $\mathbf{A}$ is denoted by $\mathbf{A}_{*j}$. Similarly, the ith row vector of $\mathbf{A}$ is denoted by $\mathbf{A}_{i*}$. The minimal and maximal singular values of $\mathbf{A}$ is denoted by $\sigma_{\min}(\mathbf{A})$ and $\sigma_{\max}(\mathbf{A})$. $\mathbf{I}_p$ denotes the $p \times p$ identity matrix. $\langle \cdot, \cdot \rangle : \mathbb{R}^{n \times m} \times \mathbb{R}^{n \times m} \to \mathbb{R}$ denotes the trace inner product given by $\langle \mathbf{A}, \mathbf{B} \rangle = \mathrm{tr}(\mathbf{A}\mathbf{B}^\top)$. For a function $f : \mathbb{R}^p \to \mathbb{R}$, and $\boldsymbol{Z}_i \in \mathbb{R}^p$, define the *empirical process* $\mathbb{G}_n(f(\boldsymbol{Z}_i)) = n^{-1/2} \sum_{i=1}^n \{f(\boldsymbol{Z}_i) - \mathsf{E}[f(\boldsymbol{Z}_i)]\}$.

Definition 4.1.1 (Sub-Gaussian variable and sub-Gaussian norm)**.** A random variable X is called *sub-Gaussian* if there exists some positive constant K_2 such that $\mathrm{P}(|X| > t) \leq \exp(1 - t^2/K_2^2)$ for all $t \geq 0$. The *sub-Gaussian norm* of X is defined as $\|X\|_{\psi_2} = \sup_{p \geq 1} p^{-1/2} (\mathsf{E}|X|^p)^{1/p}$.

4.2 Factorizable sparse multivariate quantile regression

To motivate the estimation of factors in the quantile of a random variable, we first shortly review classical linear factor model. Linear factor models, such as Capital Asset Pricing Model (CAPM) and Arbitrage Pricing Theory (APT), are popular in economics and finance for describing the relationship between asset returns and factors. The standard setting is

$$Y_{ij} = \psi_{j1} F_{i1} + \psi_{j2} F_{i2} + ... + \psi_{jr} F_{ir} + \varepsilon_{ij}, \tag{4.2.1}$$

where $\boldsymbol{Y}_i \in \mathbb{R}^m$ is a vector of asset returns, F_{i1},...,F_{ir} are factors and ε_{ij} is the portion not related to the factors. Assumptions are $\mathrm{Cov}(F_{ik}, \varepsilon_{ij}) = 0$ for all $k = 1, ..., r$ and $j = 1, ..., m$, $\mathrm{Cov}(\varepsilon_{ij}, \varepsilon_{il}) = 0$ for all $j \neq l$. Factors F_{ik} can be viewed as hedging portfolios or macroeconomic drivers depending on the context. Note that the number of factor is exactly one in terms of CAPM.

The linear factor model (4.2.1) can be estimated even when the factors are not identified ex-ante. The multivariate regression model can estimate the factors and

loadings, if it is known that some exogenous macroeconomic variables $\boldsymbol{X}_i \in \mathbb{R}^p$ are relevant to F_{ik}. Taking conditional expectation to factor model (4.2.1) gives

$$\mathsf{E}[Y_{ij}|\boldsymbol{X}_i] = \sum_{k=1}^{r} \psi_{jk}\mathsf{E}[F_{ik}|\boldsymbol{X}_i], \tag{4.2.2}$$

Suppose that $\mathsf{E}[F_{i,k}|\boldsymbol{X}_i] = \boldsymbol{\varphi}_k^\top \boldsymbol{X}_i$, where $\boldsymbol{\varphi}_k = (\varphi_{k1}, ..., \varphi_{kp})$. We have the multivariate regression model

$$\mathsf{E}[\boldsymbol{Y}_i|\boldsymbol{X}_i] = \boldsymbol{\Gamma}_{*j}^\top \boldsymbol{X}_i, \tag{4.2.3}$$

where $\boldsymbol{\Gamma}_{*j} = (\sum_{k=1}^r \psi_{j,k}\varphi_{k,1}, ..., \sum_{k=1}^r \psi_{j,k}\varphi_{k,p})$. $\boldsymbol{\Gamma}$ can be estimated with a multivariate regression model (4.1.3) with the rank of $\boldsymbol{\Gamma}$ being r. The benefit of considering such model is that this incorporates the cross-sectional information in $\boldsymbol{Y}_i$. This is closely related to multi-task learning paradigm in machine learning literature. Gibbons and Ferson (1985) was the first to present the model (4.2.3). One can also see Chapter 8 of Reinsel and Velu (1998) for detail. One remark is that for the traditional multivariate regression technique introduced in Reinsel and Velu (1998), the number of factor r is assumed to be known or has to be obtained via other method. However, using the modern regularization method of Yuan et al. (2007), Bunea et al. (2011) or Negahban and Wainwright (2011), knowing r is not necessary for estimation.

One remark is that knowing $\boldsymbol{\Gamma}$ does not trivially yield the estimate for factors and factor loadings, because the decomposition of $\boldsymbol{\Gamma} = \boldsymbol{\Phi}\boldsymbol{\Psi}$ is not unique, in which $\boldsymbol{\Phi}$ corresponds to the factors and $\boldsymbol{\Psi}$ corresponds to the factor loadings. The ideal decomposition requires $\boldsymbol{\Phi}$ to be a matrix with r nonzero columns, so that we have r factors, and $\boldsymbol{\Psi}$ is a unitary matrix. As pointed out in Section 2 of Yuan et al. (2007), this can be done via singular value decomposition. Suppose the singular value decomposition of $\boldsymbol{\Gamma}$ is $\boldsymbol{\Gamma} = \mathbf{U}\mathbf{D}\mathbf{V}^\top$, where $\mathbf{U}$ and $\mathbf{V}$ are unitary matrices and $\mathbf{D}$ is rectangular diagonal matrix with kth diagonal element being the singular value σ_k, and $\sigma_k = 0$ for $k > r$. The factor loadings $\boldsymbol{\psi}_j = \mathbf{V}_{j*}$ satisfies $\|\boldsymbol{\psi}_j\|_2 = 1$ for $1 \le j \le m$. Letting $\boldsymbol{\Phi} = \mathbf{D}^\top\mathbf{U}^\top$. $\boldsymbol{\Phi}$ has only r nonzero rows. The factor is formed as $F_{ik} = \sigma_k \mathbf{U}_{*k}\boldsymbol{X}_i$.

Conditional quantile is of our focus. We estimate the quantile of response variables Y_{ij}, $j = 1, ..., m$ parametrically as (4.2.3). Let $q_j(\tau|\boldsymbol{X}_i)$ be the conditional quantile of Y_{ij} conditional on $\boldsymbol{X}_i \in \mathbb{R}^p$, for $j = 1, ..., m$ and $i = 1, ..., n$,

$$q_j(\tau|\boldsymbol{X}_i) = \boldsymbol{X}_i^\top \boldsymbol{\Gamma}_{*j}(\tau), \tag{4.2.4}$$

where $\boldsymbol{\Gamma}_{*j}$ is jth column of matrix $\boldsymbol{\Gamma} \in \mathbb{R}^{p\times m}$, which is assumed of low rank $r \ll \min\{p, m\}$. The model is posed in a high-dimensional setting: $p, m \to \infty$ while the sample size $n \to \infty$.

Furthermore, model (4.2.4) is *factorisable*. Suppose the SVD of $\boldsymbol{\Gamma}$ is $\boldsymbol{\Gamma} = \mathbf{U}\mathbf{D}\mathbf{V}^\top$ and the number of nonzero singular values is r, similarly to (4.2.2),

$$q_j(\tau|\boldsymbol{X}_i) = \sum_{k=1}^{r} V_{j,k} f_k^\tau(\boldsymbol{X}_i), \tag{4.2.5}$$

where $f_k^\tau(\boldsymbol{X}_i) = \sigma_k \mathbf{U}_{*k}^\top \boldsymbol{X}_i$. With slight abuse of terminology, we also call $f_k^\tau(\boldsymbol{X}_i)$ "factors" with $V_{j,k}$ being "factor loadings". For mean regression (4.2.3), factorisation would give a factor model (4.2.1). In the practice of multi-task or multivariate quantile regression, factors are handy for classification and prediction. We will explore its power with real data in Section 4.7.

To find an estimator $\widehat{\mathbf{\Gamma}}$ for $\mathbf{\Gamma}$, quantile regression proposed by Koenker and Bassett (1978) allows to recover the conditional quantile of a univariate response. Our loss function

$$\widehat{\mathbf{\Gamma}}_\lambda(\tau) \stackrel{\text{def}}{=} \arg\min_{\mathbf{S}\in\mathbb{R}^{p\times m}} \left\{ (mn)^{-1} \sum_{i=1}^{n}\sum_{j=1}^{m} \rho_\tau\big(Y_{ij} - \boldsymbol{X}_i^\top \mathbf{S}_{*j}\big) + \lambda\|\mathbf{S}\|_* \right\}, \tag{4.2.6}$$

where $\rho_\tau(u) = u\{\tau - \mathbf{1}(u \le 0)\}$. The first term controls the quality of fitting, which is similar to the loss function proposed in Koenker and Portnoy (1990). The second term nuclear norm regularization is applied to encourage the accurate estimation, as the rank of the matrix $\mathbf{\Gamma}$ is degenerate and is sparse. The quantity τ is considered fixed in our discussion.

Note that $\rho_\tau(u)$ is not globally differentiable, where $0 < \tau < 1$ is a given quantile level. The idea of solving (4.2.6) is first smoothing the loss function by the method of Nesterov (2005), and then applying the fast iterative proximal gradient algorithm of Beck and Teboulle (2009). It will be shown in Theorem 4.3.3 that our method achieves the efficiency of $\mathcal{O}(1/\epsilon)$, where ϵ is a given rate of accuracy, say 10^{-6}. Nonasymptotic oracle properties of $\widehat{\mathbf{\Gamma}}$ are in Section 4.4.

4.3 Estimation

In this section, we study the estimation procedure for solving (4.2.6). The procedure of estimation is summarized in Algorithm 1. The main result on efficiency of the algorithm is Theorem 4.3.3.

The problem of solving optimization like (4.1.4) and (4.2.6) has received a lot of attention recently. One strand of literature using the proximal gradient approach, exploits the fact that the proximity operator of nuclear norm has a closed form, which performs soft-thresholding of the singular values of the input matrix. Such algorithm requires singular value decomposition (SVD) in each iteration, and this may be computationally expensive when the matrix is large. Ji and Ye (2009) and Toh and Yun (2010) propose algorithms in this line which obtain ϵ-accurate solution in $\mathcal{O}(1/\sqrt{\epsilon})$ steps. A second strand of literature reformulates the optimization problem into a semidefinite program and then applies available solvers. Though traditional solvers such as SDPT3 or SeDuMi are not suitable for high-dimensional data, Jaggi and Sulovský (2010) constructed an algorithm based on the algorithm of Hazan (2008) and applied it on large datasets. This approach avoids performing SVD in each step, but in general it requires $\mathcal{O}(1/\epsilon)$ steps to reach a ϵ-accurate solution.

Our algorithm follows the first line of proximal gradient algorithm, as the in Jaggi and Sulovský (2010) it is required that the loss function has to be differentiable. In

our simulation study we show that our algorithm is able to handle matrices with hundreds of rows and columns.

A remarkable difference between our problem to those studied in the articles mentioned above is that, beside the nuclear norm penalty term, the first term in our loss function in (4.2.6) is non-smooth, and this suggests that the direct application of proximal gradient algorithm may not generate desirable result. Therefore, there are two important questions one needs to answer: how to transform the problem so that it produces favorable properties and what is the price for such transformation? In what follows we will answer both questions by showing a procedure to smooth the non-smooth loss function and obtain the convergence rate of our algorithm. Our approach is inspired by Chen et al. (2012), which deal with sparse regression problem with non-smooth structured sparsity-inducing penalties. They apply the method of Nesterov (2005), who suggests a systematic way to approximate the non-smooth objective function by a function with Lipschitz continuous gradient, our smoothing method is based on this idea as well.

Recall that our goal is to minimize the following loss function:

$$L(\mathbf{\Gamma}) = (mn)^{-1} \sum_{i=1}^{n} \sum_{j=1}^{m} \rho_\tau \left(Y_{ij} - \boldsymbol{X}_i^\top \mathbf{\Gamma}_{*j} \right) + \lambda \|\mathbf{\Gamma}\|_* \stackrel{\text{def}}{=} \widehat{Q}_\tau(\mathbf{\Gamma}) + \lambda \|\mathbf{\Gamma}\|_*, \tag{4.3.1}$$

where $\rho_\tau(u) = u\{\tau - \mathbf{1}(u \leq 0)\}$ with given $0 < \tau < 1$.

$\widehat{Q}_\tau(\mathbf{\Gamma})$ is clearly non-smooth. To handle this problem, we introduce the dual variables Θ_{ij} to rewrite as

$$\widehat{Q}_\tau(\mathbf{\Gamma}) = \max_{\Theta_{ij} \in [\tau - 1, \tau]} (mn)^{-1} \sum_{i=1}^{n} \sum_{j=1}^{m} \Theta_{ij} \left(Y_{ij} - \boldsymbol{X}_i^\top \mathbf{\Gamma}_{*j} \right). \tag{4.3.2}$$

To see that this equation holds, note that for each pair of i, j, when $Y_{ij} - \boldsymbol{X}_i^\top \mathbf{\Gamma}_{*j} > 0$, $\Theta_{ij} = \tau$ since τ is the largest "positive" value in the interval $[\tau - 1, \tau]$; when $Y_{ij} - \boldsymbol{X}_i^\top \mathbf{\Gamma}_{*j} \leq 0$, $\Theta_{ij} = \tau - 1$ since τ is the smallest "negative" value in the interval $[\tau - 1, \tau]$. This verifies the equation. Observe that it is necessary to choose $[\tau - 1, \tau]$ rather than $\{\tau - 1, \tau\}$ for the support of Θ_{ij} in order to satisfy the conditions given in Nesterov (2005). Though both choices fulfill the equation, the previous one is an interval and therefore a convex set while the later one is not convex. This choice is the key to the smoothing approximation discussed later and will influence the gradient of the smoothed loss function.

The formulation of $\widehat{Q}_\tau(\mathbf{\Gamma})$ given in (4.3.2) is still a non-smooth function of $\mathbf{\Gamma}$, and this makes the subgradient based algorithm inefficient. To smooth this function, denote $\mathbf{\Theta} = (\Theta_{ij})$ the matrix of Θ_{ij}, we consider the smooth approximation to $\widehat{Q}_\tau(\mathbf{\Gamma})$:

$$\widehat{Q}_{\tau,\kappa}(\mathbf{\Gamma}) = \max_{\Theta_{ij} \in [\tau - 1, \tau]} \left\{ (mn)^{-1} \ell(\mathbf{\Gamma}, \mathbf{\Theta}) - \frac{\kappa}{2} \|\mathbf{\Theta}\|_{\mathrm{F}}^2 \right\}, \tag{4.3.3}$$

where $\ell(\mathbf{\Gamma}, \mathbf{\Theta}) = \sum_{i=1}^{n} \sum_{j=1}^{m} \Theta_{ij} \left(Y_{ij} - \boldsymbol{X}_i^\top \mathbf{\Gamma}_{*j} \right)$, and $\kappa > 0$ is a smoothing regularization constant depending on m, n and the desired accuracy. When $\kappa \to 0$, the approximation is getting closer to the function before smoothing. We anlayse the convergence rate of our algorithm based on Theorem 1 of Nesterov (2005).

LEMMA 4.3.1. $\ell(\mathbf{\Gamma}, \mathbf{\Theta})$ can be expressed as $\ell(\mathbf{\Gamma}, \mathbf{\Theta}) = \langle -\mathbf{X}\mathbf{\Gamma}, \mathbf{\Theta}\rangle + \langle \mathbf{Y}, \mathbf{\Theta}\rangle$.

Since the function $\frac{\kappa}{2}\|\mathbf{\Theta}\|_{\mathrm{F}}^2$ is strongly convex, the optimal solution $\mathbf{\Theta}^*(\mathbf{\Gamma})$ for achieving (4.3.3) is unique for each $\mathbf{\Gamma}$. We introduce a notation: for any matrix $\mathbf{A} = (A_{ij})$, $[[\mathbf{A}]]_\tau = ([[A_{ij}]]_\tau)$ where

$$[[A_{ij}]]_\tau = \begin{cases} \tau, & \text{if } A_{ij} \geq \tau; \\ A_{ij}, & \text{if } \tau - 1 < A_{ij} < \tau; \\ \tau - 1, & \text{if } A_{ij} \leq \tau - 1. \end{cases}$$

This function performs componentwise projection on a real matrix to the interval $[\tau - 1, \tau]$. The next theorem presents properties of the (smooth) function $\widehat{Q}_{\tau,\kappa}(\mathbf{\Gamma})$.

THEOREM 4.3.2. For any $\kappa > 0$, $\widehat{Q}_{\tau,\kappa}(\mathbf{\Gamma})$ is well-defined, convex and continuously-differentiable function in $\mathbf{\Gamma}$ with the gradient $\nabla\widehat{Q}_{\tau,\kappa}(\mathbf{\Gamma}) = -(mn)^{-1}\mathbf{X}^\top\mathbf{\Theta}^*(\mathbf{\Gamma}) \in \mathbb{R}^{p\times m}$, where $\mathbf{\Theta}^*(\mathbf{\Gamma})$ is the optimal solution to (4.3.3), namely

$$\mathbf{\Theta}^*(\mathbf{\Gamma}) = [[(\kappa mn)^{-1}(\mathbf{Y} - \mathbf{X}\mathbf{\Gamma})]]_\tau. \tag{4.3.4}$$

The gradient $\nabla\widehat{Q}_{\tau,\kappa}(\mathbf{\Gamma})$ is Lipschitz continuous with the Lipschitz constant $M = (\kappa m^2 n^2)^{-1}\|\mathbf{X}\|^2$.

By inserting (4.3.4) into the equation of $\nabla\widehat{Q}_{\tau,\kappa}(\mathbf{\Gamma})$, we arrive at the gradient which will be applied in our algorithm:

$$\nabla\widehat{Q}_{\tau,\kappa}(\mathbf{\Gamma}) = -(mn)^{-1}\mathbf{X}^\top[[(\kappa mn)^{-1}(\mathbf{Y} - \mathbf{X}\mathbf{\Gamma})]]_\tau. \tag{4.3.5}$$

Observe that (4.3.5) is similar to the subgradient $-\mathbf{X}\{\tau - \mathbf{1}(\mathbf{Y} - \mathbf{X}\mathbf{\Gamma} \leq 0)\}$ of $\widehat{Q}_\tau(\mathbf{\Gamma})$, where the operator $\tau - \mathbf{1}(\cdot \leq 0)$ applies componentwise to the matrix $\mathbf{Y} - \mathbf{X}\mathbf{\Gamma}$ with a slight abuse of notation. The major difference lies in the fact that (4.3.5) replaces the discrete non-Lipschitz $\tau - \mathbf{1}(\mathbf{Y} - \mathbf{X}\mathbf{\Gamma} \leq 0)$ with a Lipschitz function $[[\kappa^{-1}(\mathbf{Y} - \mathbf{X}\mathbf{\Gamma})]]_\tau$. Figure 4.3.1 illustrates this approximation property in a univariate framework with $m = n = 1$ and $\mathbf{X} = 1$. Denote $\psi_\tau(u) = \tau - \mathbf{1}(u \leq 0)$ the subgradient of $\rho_\tau(u)$. The solid line pictures the function $\psi_\tau(u)$ with $\tau = 0.5$, which has a jump at the origin. The dashed line corresponding to the smoothing approximation gradient $[[\kappa^{-1}(\mathbf{Y} - \mathbf{X}\mathbf{\Gamma})]]_\tau$ associated with $\kappa = 0.5$, which connects the discontinuous part and joins the function $\psi_\tau(u)$ when it reaches τ the right end and $\tau - 1$ at the left end. As κ decreases to 0.05, we observe that the smoothing approximation function is getting steeper around the origin and closer to ρ_τ.

Let $S_\lambda(\cdot)$ be the proximity operator given in Theorem C.3.2. We state the main result in this section Algorithm 1 for the optimization problem (4.2.6).

The convergence rate of Algorithm 1 is given by the following theorem.

THEOREM 4.3.3 (Convergence analysis of Algorithm 1)**.** Let $\{\mathbf{\Gamma}_t\}_{t=0}^T$ be the sequence generated by Algorithm 1, and $\mathbf{\Gamma}^*$ be the optimal solution for minimizing (4.3.1). Let $\mu(\tau) = \max\{\tau, 1 - \tau\}$. Then for any t and $\epsilon > 0$,

$$|L(\mathbf{\Gamma}_t) - L(\mathbf{\Gamma}^*)| \leq \frac{\epsilon\mu(\tau)^2}{2} + \frac{4mn\|\mathbf{\Gamma}_0 - \mathbf{\Gamma}^*\|_{\mathrm{F}}^2\|\mathbf{X}\|^2}{(t+1)^2\epsilon}. \tag{4.3.6}$$

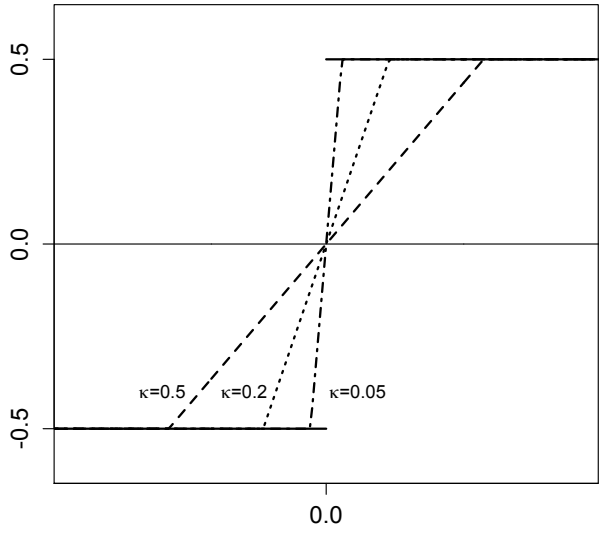


Figure 4.3.1: The solid line is the function $\psi_\tau(u) = \tau - \mathbf{1}(u \leq 0)$ with $\tau = 0.5$, which has a jump at the origin. The dashed line corresponding to the smoothing gradient $[[\kappa^{-1}(\mathbf{Y} - \mathbf{X}\boldsymbol{\Gamma})]]_\tau$ associated with $\kappa = 0.5$. As κ decreases to 0.05, we observe that the smoothing approximation function is closer to $\psi_\tau(u)$.

Algorithm 1: Smoothing fast iterative shrinkage-thresholding algorithm (SFISTA)

Input: $\mathbf{Y}$, $\mathbf{X}$, λ, $\kappa = \frac{\epsilon}{2mn}$, $M = \frac{1}{\kappa m^2 n^2}\|\mathbf{X}\|^2$;
Initialization: $\boldsymbol{\Gamma}_0 = 0$, $\boldsymbol{\Omega}_1 = 0$, step size $\delta_1 = 1$;
for $t = 1, 2, ..., T$ **do**
  $\boldsymbol{\Gamma}_t = S_{\lambda/M}\left(\boldsymbol{\Omega}_t - \frac{1}{M}\nabla\widehat{Q}_{\tau,\kappa}(\boldsymbol{\Omega}_t)\right)$;
  $\delta_{t+1} = \frac{1+\sqrt{1+4\delta_t^2}}{2}$;
  $\boldsymbol{\Omega}_{t+1} = \boldsymbol{\Gamma}_t + \frac{\delta_t - 1}{\delta_{t+1}}(\boldsymbol{\Gamma}_t - \boldsymbol{\Gamma}_{t-1})$;
end
Output $\widehat{\boldsymbol{\Gamma}} = \boldsymbol{\Gamma}_T$

If we require $L(\boldsymbol{\Gamma}_t) - L(\boldsymbol{\Gamma}^*) \leq \epsilon$, then

$$t \geq 2\frac{\sqrt{mn}\|\boldsymbol{\Gamma}^* - \boldsymbol{\Gamma}_0\|_{\mathrm{F}}\|\mathbf{X}\|}{\epsilon\left(1 - \frac{\mu(\tau)^2}{2}\right)}. \tag{4.3.7}$$

REMARK 4.3.4. 1. The first term on the right hand side of (4.3.6) is related to the smoothing error, which cannot be made small by increasing the number of iteration, but can only be reduced by choosing a smaller smoothing parameter

κ. This is the price we pay for the smoothness. The second term is related to the fast iterative proximal gradient algorithm of Beck and Teboulle (2009).

2. The original FISTA algorithm without smoothing yield the convergence rate $\mathcal{O}(1/\sqrt{\epsilon})$. In our case, smoothing approximation error deteriorates the convergence rate and the best we can do is $\mathcal{O}(1/\epsilon)$, which is comparable to the rate obtained by Nesterov (2005). As an improvement, our rate is still better than $\mathcal{O}(1/\epsilon^2)$ given by the general subgradient method.

3. The quantile level τ enters the numerical bound (4.3.6) by a factor $\left(1-\frac{\mu(\tau)^2}{2}\right)^{-1}$, which increases when τ is getting close to the boundary of $(0,1)$.

4.4 Oracle inequalities

In this section we present the non-asymptotic oracle bounds of the estimator $\widehat{\mathbf{\Gamma}}$ defined in (4.2.6). The main results are Theorem 4.4.4 and Corollary 4.4.6, which are established through the convexity and geometric argument of Belloni and Chernozhukov (2011), concentration inequalities, and $\mathcal{E}$-net arguments.

Our risk bounds resemble the corresponding results of multivariate regression for mean, such as those in Negahban and Wainwright (2011) and Koltchinskii et al. (2011). We will compare our results to theirs in Remark 4.4.7. Koltchinskii (2013) presents an oracle inequality for excess on nuclear norm penalized convex empirical risk minimization. We cannot apply their result because our quantile loss function is not differentiable. In a novel paper, Belloni and Chernozhukov (2011) develop theory for high-dimensional Lasso estimator of non-multivariate regression for quantiles. The idea to prove their main theorem is very general and can be adapted to our case of multivariate regression for quantiles. However, some technical properties still need to be established before their method can be applied.

Let $(\boldsymbol{X}_1, \boldsymbol{Y}_1), ..., (\boldsymbol{X}_n, \boldsymbol{Y}_n)$ be i.i.d. copies of $(\boldsymbol{X}, \boldsymbol{Y})$ random vectors in $\mathbb{R}^{p+m}$. Recall $\rho(u) = \big|\mathbf{1}(u \le 0) - \tau\big||u|$ and its subgradient $\psi(u) = \tau - \mathbf{1}(u \le 0)$, and that $\widehat{\mathbf{\Gamma}}$ is defined as (4.2.6). Recall also the empirical loss

$$\widehat{Q}_\tau(\mathbf{S}) = (nm)^{-1} \sum_{i=1}^{n} \sum_{j=1}^{m} \rho\left(Y_{ij} - \boldsymbol{X}_i^\top \mathbf{S}_{*j}\right)$$

and its expectation $Q_\tau(\mathbf{S})$. We define $\mathbf{\Gamma}$ be the minimizer of $Q_\tau(\mathbf{S})$, and the difference $\widehat{\mathbf{\Delta}} = \widehat{\mathbf{\Gamma}} - \mathbf{\Gamma}$. The subgradient for the empirical loss function $\widehat{Q}_\tau(\mathbf{\Gamma})$ is the matrix

$$\nabla \widehat{Q}_\tau(\mathbf{\Gamma}) = (nm)^{-1} \sum_{i=1}^{n} \boldsymbol{X}_i \boldsymbol{W}_i^\top = (nm)^{-1} \mathbf{X}^\top \mathbf{W} \in \mathbb{R}^{p \times m},$$

where $\mathbf{X}$ is the design matrix and

$$\boldsymbol{W}_i \overset{\text{def}}{=} \left(\mathbf{1}(Y_{ij} - \boldsymbol{X}_i^\top \mathbf{\Gamma}_{*j} \le 0) - \tau\right)_{1 \le j \le m}, \quad \mathbf{W} = [\boldsymbol{W}_1, ..., \boldsymbol{W}_n]^\top \in \mathbb{R}^{n \times m}.$$

In what follows we generalize the support of high-dimensional vector recovery to matrix by projections. If $\mathbf{A} \in \mathbb{R}^{p\times m}$ is of rank r, and the singular value decomposition of $\mathbf{A}$ is $\mathbf{A} = \sum_{j=1}^{r} \sigma(\mathbf{A})\mathbf{u}_j\mathbf{v}_j^\top$ with orthogonal vectors $\mathbf{u}_1, ..., \mathbf{u}_r \in \mathbb{R}^p$ and $\mathbf{v}_1, ..., \mathbf{v}_r \in \mathbb{R}^m$, the *support* of $\mathbf{A}$ is defined by (S_1, S_2) in which $S_1 = \text{span}\{\mathbf{u}_1, ..., \mathbf{u}_r\}$ and $S_2 = \text{span}\{\mathbf{v}_1, ..., \mathbf{v}_r\}$. We define the projection matrix on S_1 by $\mathbf{P}_1 = \mathbf{U}_r(\mathbf{U}_r^\top\mathbf{U}_r)^{-1}\mathbf{U}_r^\top = \mathbf{U}_r\mathbf{U}_r^\top$ in which $\mathbf{U}_r$ is a $p \times r$ matrix whose columns are formed by $\{\mathbf{u}_1, ..., \mathbf{u}_r\}$, and $\mathbf{U}_r^\top\mathbf{U}_r = \mathbf{I}_r$ because $\{\mathbf{u}_1, ..., \mathbf{u}_r\}$ is an orthonormal basis. Similarly, $\mathbf{P}_2 = \mathbf{V}_r\mathbf{V}_r^\top$. On the other hand, define the orthogonal projection of $\mathbf{P}_1$ and $\mathbf{P}_2$ by $\mathbf{P}_1^\perp$ and $\mathbf{P}_2^\perp$. For any matrix $\mathbf{S} \in \mathbb{R}^{p\times m}$, we define the projections:

$$\mathcal{P}_{\mathbf{A}}(\mathbf{S}) \stackrel{\text{def}}{=} \mathbf{S} - \mathbf{P}_1^\perp\mathbf{S}\mathbf{P}_2^\perp; \quad \mathcal{P}_{\mathbf{A}}^\perp(\mathbf{S}) \stackrel{\text{def}}{=} \mathbf{P}_1^\perp\mathbf{S}\mathbf{P}_2^\perp.$$

Define the cone

$$\mathcal{K}(\mathbf{\Gamma}; c_0) \stackrel{\text{def}}{=} \left\{\mathbf{S} \in \mathbb{R}^{p\times m} : \|\mathcal{P}_{\mathbf{A}}^\perp(\mathbf{S})\|_* \le c_0\|\mathcal{P}_{\mathbf{A}}(\mathbf{S})\|_*\right\}. \tag{4.4.1}$$

Assumption 4.1 (Sampling setting)**.** Samples $(\boldsymbol{X}_1, \boldsymbol{Y}_1), ..., (\boldsymbol{X}_n, \boldsymbol{Y}_n)$ are i.i.d. copies of $(\boldsymbol{X}, \boldsymbol{Y})$ random vectors in $\mathbb{R}^{p+m}$. $F^{-1}_{Y_{ij}|\boldsymbol{X}_i}(\tau|\boldsymbol{x}) = \boldsymbol{x}^\top\mathbf{\Gamma}_{*j}(\tau)$. Conditioning on $\boldsymbol{X}_i$, Y_{ij} is independent in j.

Assumption 4.1 postulates that the data are i.i.d and there is no cross-sectional dependence in $Y_{i1}, ..., Y_{1m}$ once conditioning on $\boldsymbol{X}_i$. This suggests that all dependency in the components of $\boldsymbol{Y}_i$ is captured by the covariates $\boldsymbol{X}_i$. This assumption is stronger than that usually required for factor models, for which uncorrelatedness is often sufficient.

Assumption 4.2 (Covariance matrix condition)**.** Let the covariance matrix of $\boldsymbol{X}$ be $\mathbf{\Sigma}_{\boldsymbol{X}}$, assume that $0 < \sigma_{\min}(\mathbf{\Sigma}_{\boldsymbol{X}}) < \sigma_{\max}(\mathbf{\Sigma}_{\boldsymbol{X}}) < \infty$. Moreover, assume the sample covariance matrix of covariates $\widehat{\mathbf{\Sigma}}_{\boldsymbol{X}} = \frac{1}{n}\mathbf{X}^\top\mathbf{X}$ satisfies

$$\mathrm{P}\big[\sigma_{\min}(\widehat{\mathbf{\Sigma}}_{\boldsymbol{X}}) \ge c_1\sigma_{\min}(\mathbf{\Sigma}_{\boldsymbol{X}}), \sigma_{\max}(\widehat{\mathbf{\Sigma}}_{\boldsymbol{X}}) \le c_2\sigma_{\max}(\mathbf{\Sigma}_{\boldsymbol{X}})\big] \ge 1 - \gamma_n. \tag{4.4.2}$$

When the covariates come from a joint p-Gaussian distribution $N(0, \mathbf{\Sigma}_{\boldsymbol{X}})$, Lemma C.3.5 shows that (4.4.2) holds with $c_1 = 1/9$, $c_2 = 9$ and $\gamma_n = 4\exp(-n/2)$.

Assumption 4.3 (Conditional density condition)**.** There exist $\underline{f} > 0$ and $\bar{f}' < \infty$ such that $|\frac{\partial}{\partial y_j} f_{Y_{ij}|\boldsymbol{X}_i}(y_i|\boldsymbol{x})| \le \bar{f}'$ and $\inf_{j\le m}\inf_{\boldsymbol{x}} f_{Y_{ij}|\boldsymbol{X}_i}(\boldsymbol{x}^\top\mathbf{\Gamma}_{*j}|\boldsymbol{x}) \ge \underline{f}$, where $f_{Y_{ij}|\boldsymbol{X}_i}$ is the conditional density function of Y_{ij} on $\boldsymbol{X}_i$.

Similar condition as Assumption 4.3 is also found in Belloni and Chernozhukov (2011). The quantity $\underline{f}$ controls the curvature of the population loss function, which can influence the estimation error. Negahban et al. (2012) give an extensive account on this issue.

Assumption 4.4 (Restricted eigenvalue and nonlinearity)**.** For a given probability distribution Π for $\boldsymbol{X}$,

$$\beta_{\boldsymbol{\Gamma},3} \stackrel{\text{def}}{=} \inf\left\{\beta > 0 : \beta\|\mathcal{P}_{\boldsymbol{\Gamma}}(\boldsymbol{\Delta})\|_{\mathrm{F}} \le \|\boldsymbol{\Delta}\|_{L_2(\Pi)},\ \forall \boldsymbol{\Delta} \in \mathcal{K}(\boldsymbol{\Gamma},3)\right\} > 0, \tag{4.4.3}$$

$$\nu \stackrel{\text{def}}{=} \frac{3}{8}\frac{\underline{f}}{\overline{f}'} \inf_{\substack{\boldsymbol{\Delta}\in\mathcal{K}(\boldsymbol{\Gamma},3)\\ \boldsymbol{\Delta}\neq 0}} \frac{\|\boldsymbol{\Delta}\|_{L_2(\Pi)}^3}{m^{-1}\sum_{j=1}^m \mathsf{E}[|\boldsymbol{X}_i^\top \boldsymbol{\Delta}_{*j}|^3]} > 0, \tag{4.4.4}$$

where $\|\mathbf{S}\|_{L_2(\Pi)}^2 \stackrel{\text{def}}{=} m^{-1}\mathsf{E}_\Pi\|\mathbf{S}^\top \boldsymbol{X}_i\|_2^2$.

The cone $\mathcal{K}(\boldsymbol{\Gamma},3)$ appears often in Lasso literature, for example in Bickel et al. (2009) and Negahban and Wainwright (2011). Similar assumption on the existence of constant $\beta_{\boldsymbol{\Gamma},3}$ can also be found in Negahban and Wainwright (2011) and Koltchinskii et al. (2011). From Assumption 4.2 and the fact that $\|\mathcal{P}_{\boldsymbol{\Gamma}}(\boldsymbol{\Delta})\|_{\mathrm{F}} \le \|\boldsymbol{\Delta}\|_{\mathrm{F}}$, we have a rough lower bound $\beta_{\boldsymbol{\Gamma},3} \ge m^{-1/2}\sqrt{\sigma_{\min}(\boldsymbol{\Sigma}_{\boldsymbol{X}})}$.

The restricted nonlinearity constant ν is proposed by Belloni and Chernozhukov (2011), which is used to control the quality of minorization given in Lemma 4.4.2 (i). Section 2.5 of Belloni and Chernozhukov (2011) calculate ν for various data generating processes under different design.

The following lemma asserts that the empirical error $\widehat{\boldsymbol{\Gamma}} - \boldsymbol{\Gamma}$ lies in the cone $\mathcal{K}(\boldsymbol{\Gamma},3)$. The proof can be found in Section C.2.1

LEMMA 4.4.1. Suppose $\lambda \ge 2\|\nabla\widehat{Q}(\boldsymbol{\Gamma})\|$ and $\widehat{\boldsymbol{\Delta}} = \widehat{\boldsymbol{\Gamma}} - \boldsymbol{\Gamma}$. Then $\|\mathcal{P}_{\boldsymbol{\Gamma}}^{\perp}(\widehat{\boldsymbol{\Delta}})\|_* \le 3\|\mathcal{P}_{\boldsymbol{\Gamma}}(\widehat{\boldsymbol{\Delta}})\|_*$. That is, $\widehat{\boldsymbol{\Delta}} \in \mathcal{K}(\boldsymbol{\Gamma},3)$.

The next lemma characterizes useful properties which will be used later. The proof can be found in Section C.2.2.

LEMMA 4.4.2. Under Assumptions 4.3 and 4.4, we have

1. If $\|\boldsymbol{\Delta}\|_{L_2(\Pi)} \le 4\nu$ and $\boldsymbol{\Delta} \in \mathcal{K}(\boldsymbol{\Gamma},3)$, $Q_\tau(\boldsymbol{\Gamma}+\boldsymbol{\Delta}) - Q_\tau(\boldsymbol{\Gamma}) \ge \frac{1}{4}\underline{f}\|\boldsymbol{\Delta}\|_{L_2(\Pi)}$;
2. If $\boldsymbol{\Delta} \in \mathcal{K}(\boldsymbol{\Gamma},3)$, $\|\boldsymbol{\Delta}\|_* \le \frac{4\sqrt{2r}}{\beta_{\boldsymbol{\Gamma},3}}\|\boldsymbol{\Delta}\|_{L_2(\Pi)}$, where $r = \operatorname{rank}(\boldsymbol{\Gamma})$.

The following technical lemma characterizes the convergence rate on the empirical error of the loss function. In the proof we repeatedly apply the Hoeffding's inequalities and Assumption 4.2. The proof can be found in Section C.2.3

LEMMA 4.4.3. Under Assumptions 4.1-4.4. Let

$$\mathcal{A}(t) = \sup_{\substack{\|\boldsymbol{\Delta}\|_{L_2(\Pi)}\le t,\\ \boldsymbol{\Delta}\in\mathcal{K}(\boldsymbol{\Gamma},3)}} \left|\mathbb{G}_n\left[m^{-1}\sum_{j=1}^m \left(\rho_\tau\{Y_{ij} - \boldsymbol{X}_i^\top(\boldsymbol{\Gamma}_{*j}+\boldsymbol{\Delta}_{*j})\} - \rho_\tau\{Y_{ij} - \boldsymbol{X}_i^\top\boldsymbol{\Gamma}_{*j}\}\right)\right]\right|. \tag{4.4.5}$$

Then

$$\mathcal{A}(t) \le \left(\sqrt{\frac{2\{\tau\vee(1-\tau)\}}{C'}} + 2\right)\frac{\alpha t\sqrt{c_2\sigma_{\max}(\boldsymbol{\Sigma}_{\boldsymbol{X}})\log(p+m)}}{m}$$

with probability greater than $1-9(p+m)^{-2}-\gamma_n$, where c_2, C' are universal constants from Assumption 4.2 and Lemma C.3.3, $\alpha = \frac{4\sqrt{2r}}{\beta_{\mathbf{\Gamma},3}}$ with $r = \text{rank}(\mathbf{\Gamma})$, $\beta_{\mathbf{\Gamma},3}$ from Assumption 4.4, and $p+m > 3$.

The following theorem derives the bounds for the prediction error, Frobenius and nuclear norm, expressed in terms of λ, condition number $\mathbf{\Sigma}_{\boldsymbol{X}}$, τ and $\underline{f}$. The proof follows similar steps as proving Theorem 2 in Belloni and Chernozhukov (2011), which explicitly exploits the convexity of the loss function and cone condition.

THEOREM 4.4.4. Under Assumptions 4.1-4.4, $\lambda \geq 2\|\nabla \widehat{Q}(\mathbf{\Gamma})\|$ and the growth condition on r:

$$\left(C_\tau \frac{\sqrt{\sigma_{\max}(\mathbf{\Sigma}_{\boldsymbol{X}})\log(p+m)}}{m\sqrt{n}\underline{f}} + \frac{\lambda}{\underline{f}}\right)\frac{4\sqrt{2r}}{\beta_{\mathbf{\Gamma},3}} < \nu. \tag{4.4.6}$$

Then

$$\|\widehat{\mathbf{\Gamma}} - \mathbf{\Gamma}\|_{L_2(\Pi)} \leq 4C_\tau \frac{\alpha\sqrt{\sigma_{\max}(\mathbf{\Sigma}_{\boldsymbol{X}})\log(p+m)}}{m\sqrt{n}\underline{f}} + 4\lambda\frac{\alpha}{\underline{f}} \tag{4.4.7}$$

$$\|\widehat{\mathbf{\Gamma}} - \mathbf{\Gamma}\|_{\text{F}} \leq 4C_\tau \frac{\alpha}{\sqrt{m}\underline{f}}\sqrt{\frac{\sigma_{\max}(\mathbf{\Sigma}_{\boldsymbol{X}})}{\sigma_{\min}(\mathbf{\Sigma}_{\boldsymbol{X}})}}\sqrt{\frac{\log(p+m)}{n}} + 4\lambda\frac{\sqrt{m}\alpha}{\sigma_{\min}(\mathbf{\Sigma}_{\boldsymbol{X}})\underline{f}} \tag{4.4.8}$$

$$\|\widehat{\mathbf{\Gamma}} - \mathbf{\Gamma}\|_{*} \leq 4C_\tau \frac{\alpha^2\sqrt{\sigma_{\max}(\mathbf{\Sigma}_{\boldsymbol{X}})}}{m\underline{f}}\sqrt{\frac{\log(p+m)}{n}} + 4\lambda\frac{\alpha^2}{\underline{f}} \tag{4.4.9}$$

with probability $1-9(p+m)^{-2}-\gamma_n$, where $\alpha = \frac{4\sqrt{2r}}{\beta_{\mathbf{\Gamma},3}}$ with $r = \text{rank}(\mathbf{\Gamma})$, $\beta_{\mathbf{\Gamma},3}$ from Assumption 4.4, $C_\tau = \left(\sqrt{\frac{2\{\tau\vee(1-\tau)\}}{C'}}+2\right)\sqrt{c_2}$, $C' > 0$ is a universal constant from Lemma C.3.3, c_2 from Assumption 4.2 and $p+m > 3$.

Proof of Theorem 4.4.4. Let

$$\begin{aligned}\Omega_1 = &\text{ the event that Assumption 4.2 holds;}\\ \Omega_2 = &\text{ the event } \mathcal{A}(t) \leq C_\tau \frac{\alpha t\sqrt{\sigma_{\max}(\mathbf{\Sigma}_{\boldsymbol{X}})\log(p+m)}}{m}.\end{aligned}$$

Note that the probability of event $\text{P}(\Omega_1\cap\Omega_2) \geq 1-\gamma_n-9(p+m)^{-2}$. Set

$$t = 4C_\tau \frac{\alpha\sqrt{\sigma_{\max}(\mathbf{\Sigma}_{\boldsymbol{X}})\log(p+m)}}{m\sqrt{n}\underline{f}} + 4\lambda\frac{\alpha}{\underline{f}} > 0.$$

We show that on $\Omega_1\cap\Omega_2$, $\|\boldsymbol{X}^\top\widehat{\mathbf{\Delta}}\| > t$ is infeasible. Let $\widehat{\mathbf{\Delta}} = \widehat{\mathbf{\Gamma}} - \mathbf{\Gamma}$. On event $\{\|\boldsymbol{X}^\top\widehat{\mathbf{\Delta}}\| \geq t\}$, from Lemma 4.4.1, one has

$$0 > \inf_{\|\mathbf{\Delta}\|_{L_2(\Pi)}\geq t, \mathbf{\Delta}\in\mathcal{K}(\mathbf{\Gamma},3)} \widehat{Q}_\tau(\mathbf{\Gamma}+\mathbf{\Delta}) - \widehat{Q}_\tau(\mathbf{\Gamma}) + \lambda(\|\mathbf{\Gamma}+\mathbf{\Delta}\|_* - \|\mathbf{\Gamma}\|_*), \tag{4.4.10}$$

As argued in the proof of Theorem 2 of Belloni and Chernozhukov (2011), from the facts that

1. The minimizer of $\widehat{Q}_\tau(\cdot) + \lambda\|\cdot\|_*$ is unique due to its convexity;

2. $\mathcal{K}(\mathbf{\Gamma}, 3)$ is a cone,

on the boundary $\{\mathbf{\Delta} \in \mathcal{K}(\mathbf{\Gamma}, 3) : \|\mathbf{\Delta}\|_{L_2(\Pi)} = t\}$ the loss function $\widehat{Q}(\mathbf{\Gamma}+\mathbf{\Delta}) + \lambda\|\mathbf{\Gamma}+\mathbf{\Delta}\|_*$ is less than that of $\widehat{Q}(\mathbf{\Gamma}) + \lambda\|\mathbf{\Gamma}\|_*$ ($\mathbf{\Delta} = 0$), no matter whether the optimal solution $\widehat{\mathbf{\Delta}} \in \{\mathbf{\Delta} \in \mathcal{K}(\mathbf{\Gamma}, 3) : \|\mathbf{\Delta}\|_{L_2(\Pi)} \geq t\}$ or $\widehat{\mathbf{\Delta}} \in \{\mathbf{\Delta} \in \mathcal{K}(\mathbf{\Gamma}, 3) : \|\mathbf{\Delta}\|_{L_2(\Pi)} < t\}$. Hence, we have the inequality

$$0 > \inf_{\|\mathbf{\Delta}\|_{L_2(\Pi)}=t, \mathbf{\Delta}\in\mathcal{K}(\mathbf{\Gamma},3)} \widehat{Q}_\tau(\mathbf{\Gamma}+\mathbf{\Delta}) - \widehat{Q}_\tau(\mathbf{\Gamma}) + \lambda(\|\mathbf{\Gamma}+\mathbf{\Delta}\|_* - \|\mathbf{\Gamma}\|_*),$$

It can be further deducted that

$$0 > \inf_{\|\mathbf{\Delta}\|_{L_2(\Pi)}=t, \mathbf{\Delta}\in\mathcal{K}(\mathbf{\Gamma},3)} Q_\tau(\mathbf{\Gamma}+\mathbf{\Delta}) - Q_\tau(\mathbf{\Gamma}) - n^{-1/2}\mathcal{A}(t) + \lambda(\|\mathbf{\Gamma}+\mathbf{\Delta}\|_* - \|\mathbf{\Gamma}\|_*),$$

By triangle inequality, $\big|\|\mathbf{\Gamma}+\mathbf{\Delta}\|_* - \|\mathbf{\Gamma}\|_*\big| \leq \|\mathbf{\Delta}\|_* \leq \alpha\|\mathbf{\Delta}\|_{L_2(\Pi)} = \alpha t$ on the set $\{\|\mathbf{\Delta}\|_{L_2(\Pi)} = t, \mathbf{\Delta} \in \mathcal{K}(\mathbf{\Gamma}, 3)\}$. Furthermore, by Lemma 4.4.3, on event $\Omega_1 \cap \Omega_2$

$$\mathcal{A}(t) \leq C_\tau \frac{t\sqrt{\sigma_{\max}(\mathbf{\Sigma}_{\boldsymbol{X}})\log(p+m)}}{m} t.$$

Therefore, on event $\Omega_1 \cap \Omega_2$, it holds that

$$0 > \inf_{\|\mathbf{\Delta}\|_{L_2(\Pi)}=t, \mathbf{\Delta}\in\mathcal{K}(\mathbf{\Gamma},3)} Q_\tau(\mathbf{\Gamma}+\mathbf{\Delta}) - Q_\tau(\mathbf{\Gamma}) - C_\tau \frac{\alpha\sqrt{\sigma_{\max}(\mathbf{\Sigma}_{\boldsymbol{X}})\log(p+m)}}{m\sqrt{n}} t - \lambda\alpha t,$$

Finally, applying Lemma 4.4.2 (i), we have

$$0 > \inf_{\|\mathbf{\Delta}\|_{L_2(\Pi)}=t, \mathbf{\Delta}\in\mathcal{K}(\mathbf{\Gamma},3)} \frac{1}{4}\underline{f}t^2 - C_\tau \frac{\alpha\sqrt{\sigma_{\max}(\mathbf{\Sigma}_{\boldsymbol{X}})\log(p+m)}}{m\sqrt{n}} t - \lambda\alpha t. \tag{4.4.11}$$

With our choice of t, (4.4.11) cannot hold. Thus, the inequality (4.4.7) holds.

The inequality (4.4.8) can be obtained by the simple observation that $\|\mathbf{\Delta}\|^2_{L_2(\Pi)} \geq (\sigma_{\min}(\mathbf{\Sigma}_{\boldsymbol{X}})/m)\|\mathbf{\Delta}\|^2_{\mathrm{F}}$.

The inequality (4.4.9) for $\|\widehat{\mathbf{\Delta}}\|_*$ follows from the fact that $\widehat{\mathbf{\Delta}} \in \mathcal{K}(\mathbf{\Gamma}, 3)$ by Lemma 4.4.1, Lemma 4.4.2 (ii) and the bound for $\|\widehat{\mathbf{\Delta}}\|_{L_2(\Pi)}$. □

Next lemma gives the bound for $\frac{1}{n}\|\mathbf{X}^\top\mathbf{W}\|$. From which we obtain a bound for $\|\nabla\widehat{Q}(\mathbf{\Gamma})\|$.

LEMMA 4.4.5. Under Assumption 4.1 and 4.2,

$$\frac{1}{n}\|\mathbf{X}^\top\mathbf{W}\| \leq C^*\sqrt{\sigma_{\max}(\mathbf{\Sigma}_{\boldsymbol{X}})\{\tau \vee (1-\tau)\}}\sqrt{\frac{p+m}{n}}, \text{ where } C^* = 4\sqrt{2\frac{c_2}{C'}\log 8} \tag{4.4.12}$$

with probability greater than $1 - 3e^{-(p+m)\log 8} - \gamma_n$, where C' and c_2 are absolute constants given by Hoeffding's inequality C.3.3 and Assumption 4.2.

Let us take the rough bound $\beta_{\mathbf{\Gamma},3} \geq m^{-1/2}\sqrt{\sigma_{\min}(\mathbf{\Sigma}_{\boldsymbol{X}})}$. Lemma 4.4.5 and Lemma 4.4.1 suggest to take

$$\lambda = 2\frac{C^*}{m}\sqrt{\sigma_{\max}(\mathbf{\Sigma}_{\boldsymbol{X}})\{\tau \vee (1-\tau)\}}\sqrt{\frac{p+m}{n}}. \tag{4.4.13}$$

By the choice (4.4.13), Theorem 4.4.4 yields the oracle rate, which we summarize in Corollary (4.4.6).

The last result in this section gives the rate of convergence under the choice of λ given in (4.4.13), which will be the guideline for simulation comparison in Section 4.6.

COROLLARY 4.4.6. Assume that Assumptions 4.1-4.4 hold and select λ as (4.4.13). Under the growth condition on r:

$$\frac{C'_\tau}{\underline{f}\sqrt{m}}\sqrt{\frac{\sigma_{\max}(\mathbf{\Sigma}_{\boldsymbol{X}})}{\sigma_{\min}(\mathbf{\Sigma}_{\boldsymbol{X}})}}\sqrt{\tau \vee (1-\tau)}\left(\sqrt{\frac{\log(p+m)}{n}} + \sqrt{\frac{p+m}{n}}\right)\sqrt{r} < \nu. \tag{4.4.14}$$

Then

$$\|\widehat{\mathbf{\Gamma}} - \mathbf{\Gamma}\|_{L_2(\Pi)} \leq \frac{C'_\tau}{\underline{f}\sqrt{m}}\sqrt{\frac{\sigma_{\max}(\mathbf{\Sigma}_{\boldsymbol{X}})}{\sigma_{\min}(\mathbf{\Sigma}_{\boldsymbol{X}})}}\sqrt{\tau \vee (1-\tau)}\sqrt{r}\left(\sqrt{\frac{\log(p+m)}{n}} + \sqrt{\frac{p+m}{n}}\right), \tag{4.4.15}$$

$$\|\widehat{\mathbf{\Gamma}} - \mathbf{\Gamma}\|_{\mathrm{F}} \leq \frac{C'_\tau}{\underline{f}}\sqrt{\frac{\sigma_{\max}(\mathbf{\Sigma}_{\boldsymbol{X}})}{\sigma^2_{\min}(\mathbf{\Sigma}_{\boldsymbol{X}})}}\sqrt{\tau \vee (1-\tau)}\sqrt{r}\left(\sqrt{\frac{\log(p+m)}{n}} + \sqrt{\frac{p+m}{n}}\right), \tag{4.4.16}$$

$$\|\widehat{\mathbf{\Gamma}} - \mathbf{\Gamma}\|_{*} \leq \frac{C''_\tau}{\underline{f}}\sqrt{\frac{\sigma_{\max}(\mathbf{\Sigma}_{\boldsymbol{X}})}{\sigma^2_{\min}(\mathbf{\Sigma}_{\boldsymbol{X}})}}\sqrt{\tau \vee (1-\tau)}r\left(\sqrt{\frac{\log(p+m)}{n}} + \sqrt{\frac{p+m}{n}}\right), \tag{4.4.17}$$

with probability greater than $1 - \gamma_n - 9(p+m)^{-2} - 3e^{-(p+m)\log 8}$ and $p + m > 3$, where

$$C'_\tau = 8\sqrt{2}\left[\left(\sqrt{\frac{2}{C'}} + \frac{2}{\sqrt{\tau \vee (1-\tau)}}\right)\sqrt{c_2} \vee 4\sqrt{2\frac{c_2}{C'}\log 8}\right], \tag{4.4.18}$$

$C''_\tau = 4\sqrt{2}C'_\tau$ with $r = \mathrm{rank}(\mathbf{\Gamma})$, $\beta_{\mathbf{\Gamma},3}$ from Assumption 4.4 and c_2 from Assumption 4.2.

Proof of Corollary 4.4.6. Let events Ω_1 and Ω_2 be defined as in the proof of Theorem 4.4.4, and

$$\Omega_3 = \text{ the event that } \frac{1}{n}\|\mathbf{X}^\top\mathbf{W}\| \leq C^*\sqrt{\|\mathbf{\Sigma}_{\boldsymbol{X}}\|\{\tau \vee (1-\tau)\}}\sqrt{\frac{p+m}{n}}.$$

Note that the probability $\mathrm{P}(\Omega_1 \cap \Omega_2 \cap \Omega_3) \geq 1 - \gamma_n - 9(p+m)^{-2} - 3e^{-(p+m)\log 8}$. On $\Omega_1 \cap \Omega_2 \cap \Omega_3$, the bounds (4.4.7), (4.4.8), (4.4.9), and (4.4.12) hold. Inserting the rate of λ in (4.4.13) and the lower bound $\beta_{\mathbf{\Gamma},3} \geq m^{-1/2}\sqrt{\sigma_{\min}(\mathbf{\Sigma}_{\boldsymbol{X}})}$ into (4.4.7), (4.4.8),and (4.4.9) yields bounds (4.4.15), (4.4.16),and (4.4.17). □

REMARK 4.4.7. 1. The restricted nonlinearity constant ν enters the bounds only through the growth condition (4.4.14) on r. This corresponds to the Lasso for quantile regression of Belloni and Chernozhukov (2011).

2. Component of the risk bounds: Corollary 4.4.6 shows that the errors are close to the estimation error given the true model. The bounds (4.4.15), (4.4.16), and (4.4.17) consist of three components: the dimensionality, covariance matrix of the covariates and conditional density of Y given $\boldsymbol{X}$. When p and m are fixed with respect to n, the errors decrease in $n^{-1/2}$. p and m are allowed to grow with n; however, they are *not* allowed to grow faster than n for sensible estimation. This phenomenon is also found in the multivariate regression for mean. Please see Negahban and Wainwright (2011), Koltchinskii et al. (2011) among others. Rank r of matrix $\mathbf{\Gamma}$ enters the bound as a factor, and $r(p+m)$ is the number of unknown parameters. The covariates can influence the bounds (4.4.15), (4.4.16), and (4.4.17) through the condition number $\frac{\sigma_{\max}(\mathbf{\Sigma}_{\boldsymbol{X}})}{\sigma_{\min}(\mathbf{\Sigma}_{\boldsymbol{X}})}$ of the covariance matrix $\mathbf{\Sigma}_{\boldsymbol{X}}$. Large condition number also introduces instability to multivariate regression for quantiles as for mean. Finally, the minimal value of densities $\underline{f}$ and the quantile level τ are related to the conditional distribution of Y_{ij} give $\boldsymbol{X}_i$ and are only seen in multivariate regression for quantiles. We show in (4.4.15), (4.4.16), and (4.4.17) that small minimal value of densities $\underline{f}$, which may result from the large support of Y_{ij}, can result in inaccurate estimation. On the other hand, the estimation at τ close to 0 or 1 is also difficult as $\tau \vee (1-\tau)$ enters as a factor to the estimation errors.

4.5 Tuning

For implementation it is crucial to appropriately select λ. In theory, one can select λ based on (4.4.13), but the value is not adaptive to the data very well. In this section we propose a way to select λ based on the "pivotal principle", which are better adaptive to the data.

Define the random variable

$$\Lambda = (nm)^{-1}\|\mathbf{X}^\top \widetilde{\mathbf{W}}\|, \tag{4.5.1}$$

where $\widetilde{W}_{ij} = \mathbf{1}(U_{ij} \leq 0) - \tau$, $\{U_{ij}\}$ for $i = 1, ..., n$ and $j = 1, ..., m$ are i.i.d. uniform (0,1) random variables, independently distributed from the input variables $\boldsymbol{X}_1, ..., \boldsymbol{X}_n$. The random variable Λ is pivotal conditioning on design $\mathbf{X}$, as it does not depend on unknown parameter $\mathbf{\Gamma}$. Notice that $(nm)^{-1}\mathbf{X}^\top\widetilde{\mathbf{W}}$ is the score $\nabla \widehat{Q}_\tau(\mathbf{\Gamma})$. Set

$$\lambda = 2 \cdot \Lambda(1-\alpha|\mathbf{X}), \tag{4.5.2}$$

where $\Lambda(1-\alpha|\mathbf{X}) \stackrel{\text{def}}{=} (1-\alpha)$-quantile of Λ conditional on $\mathbf{X}$, and c is an absolute constant. This is the pivotal principle applied in the high-dimensional quantile regression of Belloni and Chernozhukov (2011) and square-root Lasso Belloni et al. (2011).

4.6 Simulation

In this section we check the performance of the proposed method via Monte Carlo simulations and verify the oracle properties in Section 4.4. In the first set of simulation, we consider three symmetric models, which are different in terms of the degree of sparsity. In the second set of simulation, an asymmetric setting is considered with two different degree of sparsity. We consider three symmetric models with different degrees of sparsity in Section 4.6.1. Section 4.6.2 is devoted to two asymmetric models.

4.6.1 Symmetric models

We consider three models that differ in their complexity:

- Model LS (Less sparse): Set $m = p = n = 500$. In each iteration, each entry of the $p \times m$ coefficient matrix $\mathbf{\Gamma}$ is generated from a i.i.d. normal distribution. Setting the last 375 singular values of $\mathbf{\Gamma}$ to 0;
- Model MS (Moderate sparse): Generating $\mathbf{\Gamma}$ as Model LS. Setting the first 10 singular values to 30, and 0 for the rest;
- Model ES (Extremely sparse): Generating $\mathbf{\Gamma}$ as Model LS. Replacing the first singular values by 20, and 0 for the rest.

Given the $\mathbf{\Gamma}$ generated by the model above, at each iteration, we generate $\boldsymbol{X}_i$ from $N(0, \Sigma)$ with $\sigma_{ij} = 0.5^{|i-j|}$. The response variable is generated as

$$\boldsymbol{Y}_i = \mathbf{\Gamma}^\top \boldsymbol{X}_i + \boldsymbol{\varepsilon}_i, \tag{4.6.1}$$

where $\boldsymbol{\varepsilon}_i$ is a random vector in which each element is from i.i.d. standard normal distribution.

We estimate the model at quantile levels $\tau = 0.05$, 0.1, 0.2, 0.5, 0.8, 0.9, 0.95. In order to get some ideas on the solution path, we set $\lambda = (5 \times 10^{-6}, 10^{-5}, 5 \times 10^{-5}, 10^{-4})$ for comparison purpose. For reference, using the tuning technique in Section 4.5, the simulated $\lambda = (0.00477, 0.00465, 0.00438, 0.00346)$ for $\tau = 5\%$, 10%, 20%, and 50%. The λ for $\tau = 95\%, 90\%$ and 80% are the same as that of $\tau = 5\%, 10\%, 20\%$ by symmetry. The iteration run is 500.

We stop the SFISTA algorithm at step t when the difference of loss function at step $t-1$ and t is less than 10^{-6}. Moreover, considering the size of our model and the choice of κ in the simulation study of Chen et al. (2012), we directly set $\kappa = 0.0001$, rather than applying the κ given by Theorem 4.3.3.

The performance of $\widehat{\mathbf{\Gamma}}$ is measured by:

- Prediction error: $m^{-1}\|\mathbf{X}(\widehat{\mathbf{\Gamma}} - \mathbf{\Gamma})\|_{\mathrm{F}}$;
- Model selection: Frobenius error $\|\mathbf{\Gamma} - \widehat{\mathbf{\Gamma}}\|_{\mathrm{F}}$ and nuclear error $\|\mathbf{\Gamma} - \widehat{\mathbf{\Gamma}}\|_{*}$;
- Estimated number of nonzero singular values;
- Computational time.

The number of nonzero singular values is determined by the sudden drop in singular values of $\widehat{\mathbf{\Gamma}}$. If the drop from $\hat{r}$th singular value to $\hat{r}+1$th singular value is greater than a given threshold, then we record the number of nonzero singular values as $\hat{r}$. Notice that the three symmetric models only differ in sparsity. From the simulation, we can clearly see what role sparsity plays.

The results are shown as boxplots from Figure 4.6.2 to 4.6.4. Each figure consists of five rows, which presents the prediction error, Frobenius error, nuclear error, estimated number of factors and the computational time, and the columns correspond to different values of λ.

The errors as functions in τ of the three models show "V" shape. This confirms the term $\tau \vee (1-\tau)$ appeared in the oracle bounds in Theorems 4.4.6. Furthermore, the model complexity $\mathrm{rank}(\mathbf{\Gamma})$ influences the error. Among the three models, the errors are smaller in the most sparse Model ES and larger in the less sparse Model LS. This confirms the factor $\mathrm{rank}(\mathbf{\Gamma})$ appeared in the oracle bounds given in Theorems 4.4.6.

We find that the choice of λ has something to do with sparsity, though it is not directly suggested by (4.4.13). Notice that all components involved in selecting λ in (4.4.13) are equivalent for the three symmetric models. Therefore, the same λ should apply to all three models. Nonetheless, in complex model Model LS, small λ leads to small errors; while in less complex model Model ES, larger λ which leads to small errors. In addition, λ changes the way how errors depend on τ. In Model LS, the "V" shape shown in the Frobenius and nuclear deviation becomes more flat. Hence, in such model we should choose a smaller λ when the quantile at level $\tau = 0.5$ is to be estimated, and a bigger λ when the quantiles at τ close to 0 or 1 are to be estimated.

The numbers of factors selected for the three models are generally accurate. We find that for $\tau = 0.5$ the algorithm almost always makes correct selection for all the choices of λ and all the three symmetric models. For Model ES the algorithm selects the correct number of factors even for $\tau = 0.2, 0.8$ when λ is large. For other τ, particularly the extremes ones close to 0 or 1, it is more difficult to recover the true number of factors.

About the computational efficiency of our algorithm, the time required for the algorithm to converge increases with the complexity. This fact corresponds to the term $\|\mathbf{\Gamma}^* - \mathbf{\Gamma}_0\|_{\mathrm{F}}$ in inequality (4.3.7). When we look at the most sparse Model ES Figure 4.6.4, the algorithm converges in less than 80 seconds in the best case $\lambda = 10^{-5}$. For Model LS and MS, smaller choices of λ usually imply longer time for the algorithm to converge, while larger choices of λ allow the algorithm to converge in less than 250 seconds for Model LS and 100 seconds for Model MS. On the other

hand, τ has influence on the convergence time, which corresponds to the inequality (4.3.7) and the third point of Remark 4.3.4. For example, in the last row of Figure 4.6.2 and 4.6.3, the case $\tau = 0.5$ takes least time when λ is small, but this situation reverses in the most sparse Model ES.

4.6.2 Asymmetric models

To further illustrate our method, beside adjusting the level of sparsity as done in Section 4.6.1, in this section we specify asymmetric models for the conditional distribution of Y_{ij}. Let $\mathbf{\Gamma}_1$ and $\mathbf{\Gamma}_2$ be two $p \times m$ matrices of rank r_1 and r_2 with following two specifications:

- Model AES (asymmetric extremely sparse): $(r_1, r_2) = (2, 2)$;
- Model AMS (asymmetric moderately sparse): $(r_1, r_2) = (2, 10)$.

For each model, two matrices $\mathbf{\Gamma}_1$ and $\mathbf{\Gamma}_2$ are chosen:

1. Generating vectors $\{a_1, ..., a_{r_1}\}$ and $\{b_1, ..., b_{r_2}\}$ in $\mathbb{R}^p$. The components of each vector are i.i.d. uniform distributed random variables supported on $[0, 1]$;
2. Each jth column in $\mathbf{\Gamma}_1$ is $\sum_{k=1}^{r_1} \alpha_{k,j} a_k$ where $\alpha_{k,j}$ are independent random variables in $U[0, 1]$; similarly, each jth column in $\mathbf{\Gamma}_2$ is $\sum_{k=1}^{r_2} \beta_{k,j} b_k$ where $\beta_{k,j}$ are independent random variables in $U[0, 1]$.

Now we discuss the data generation. Let U_{ij} be i.i.d. uniform random variable supported on $[0, 1]$, $i = 1, ..., n$ and $j = 1, ..., 500$. Generating $\widetilde{\boldsymbol{X}}_i$ from $N(0, \Sigma)$ with $\sigma_{ij} = 0.5^{|i-j|}$ and then setting $\boldsymbol{X}_i = \Phi(\widetilde{\boldsymbol{X}}_i)$. $\boldsymbol{X}_i$ will have support $[0, 1]^p$ and be correlated according to Falk (1999). The response variables are generated by

$$Y_{ij} = \Phi^{-1}(U_{ij}) \boldsymbol{X}_i^\top \left[\mathbf{\Gamma}_{1,*j} \mathbf{1}(U_{ij} < 0.5) + \mathbf{\Gamma}_{2,*j} \mathbf{1}(U_{ij} \geq 0.5) \right], \tag{4.6.2}$$

where $\Phi(\cdot)$ is the cdf of $N(0, 1)$. $\boldsymbol{Y}_i$ is i.i.d. by construction. Notice that when conditioning on $\boldsymbol{X}_i$, the randomness comes only from U_{ij}, which is independent of $\boldsymbol{X}_i$. Hence, Y_{ij} is independent in j when conditioning on $\boldsymbol{X}_i$.

The exact conditional quantile function $q_j(\tau|\boldsymbol{x})$ of Y_{ij} on $\boldsymbol{x}$ is

$$\begin{aligned} q_j(\tau|\boldsymbol{X}_i) &= \Phi^{-1}(\tau) \boldsymbol{X}_i^\top \mathbf{\Gamma}_{1,*j}, \quad \tau < 0.5; \\ q_j(\tau|\boldsymbol{X}_i) &= \Phi^{-1}(\tau) \boldsymbol{X}_i^\top \mathbf{\Gamma}_{2,*j}, \quad \tau \geq 0.5, \end{aligned}$$

for $j = 1, ..., 500$. Note that at $\Phi^{-1}(0.5) = 0$, and therefore the coefficient matrix at $\tau = 0.5$ is 0.

Figure 4.6.1 gives an illustration of the marginal densities of Y_{ij} for $j = 1, ...500$. The left figure is associated with Model AMS in which the densities tend to be asymmetric, in the sense that they have thick right tails and thin left tails. The densities are also more disperse. The right figure is associated with Model AES, and the densities are more symmetric and less disperse.

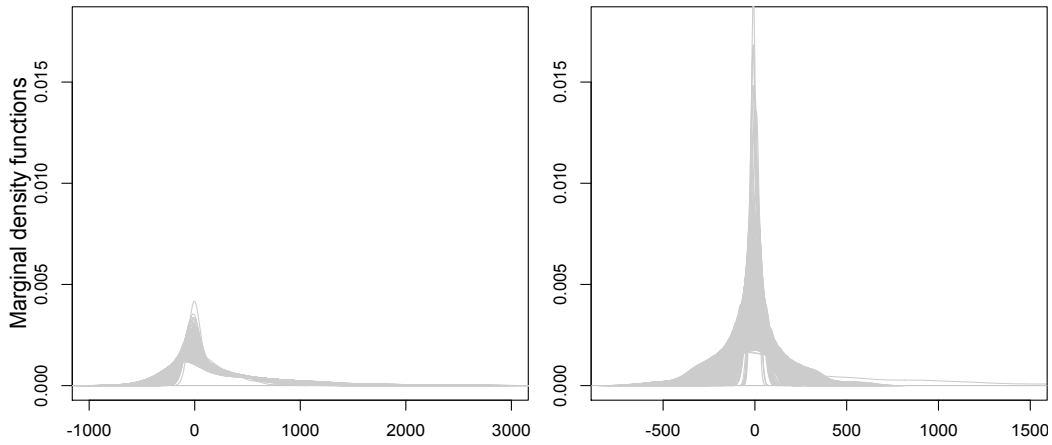


Figure 4.6.1: The plot of all 500 marginal densities of $\boldsymbol{Y}_i$ in asymmetric models. The left figure is associated with Model AMS in which the densities tend to be asymmetric (thick right tails and thin left tails). The right figure is associated with Model AES in which the densities are more symmetric.

The simulation run is 500. The measure of performance is the same as that of symmetric models. In this simulation, we select $\lambda = (0.005, 0.01, 0.05, 0.1)$. The numerical performance of the asymmetric model is shown in Figure 4.6.5 and 4.6.6. For reference, the simulated $\lambda = (0.002308, 0.002310, 0.002314, 0.002308)$ for $\tau = 5\%, 10\%, 20\%, 50\%$. The λ for $\tau = 95\%, 90\%$ and 80% are the same as that of $\tau = 5\%, 10\%, 20\%$ by symmetry.

Some common patterns can be observed from the simulated estimation errors of the two models. Firstly, the error-τ relation demonstrates a "V" shape, and this again corresponds to the factor $\tau \vee (1-\tau)$ in Theorems 4.4.6. Despite the fact that $\boldsymbol{\Gamma}_1 \neq \boldsymbol{\Gamma}_2$, the asymmetry in distribution is not significant and the error as a function of τ from Model AES is in symmetric "V" shape. In terms of the choice of λ, small λ appears to give smaller errors for both models.

The two models differ in some ways. The errors corresponding to $\tau > 0.5$ in Model AMS are notably higher than those in Model AES. This is owing to the fact that the matrix $\boldsymbol{\Gamma}_2$ in Model AMS is less sparse than Model AES. This simulation result confirms the factor $\text{rank}(\boldsymbol{\Gamma})$ in the oracle bounds in Section 4.4.

The number of nonzero singular values is almost always correctly estimated in Model AES. As expected, the estimated number of nonzero singular values of Model AMS is higher than that in Model AES when $\tau > 0.5$. However, we find that the estimated number of nonzero singular values is 2 in Model AES and between 5-7 in Model AMS, seemingly the average of the rank of $\boldsymbol{\Gamma}_1$ and $\boldsymbol{\Gamma}_2$. However, the true number of nonzero singular values at $\tau = 0.5$ is exactly 0. This shows that the singular values are hard to be accurately estimated if the coefficient matrix $\boldsymbol{\Gamma}_\tau$ is not continuous in τ.

The computational time generally follows the rule of (4.3.7). When λ is small, we find that the variation of $\tau = 0.5$ tends to be large. Due to the inflation of

rank($\mathbf{\Gamma}_2$) in Model AMS, it is more computationally demanding to recover $\widehat{\mathbf{\Gamma}}$ for $\tau > 0.5$, as implied by the term $\|\mathbf{\Gamma}^* - \mathbf{\Gamma}_0\|_{\mathrm{F}}$ in inequality (4.3.7).

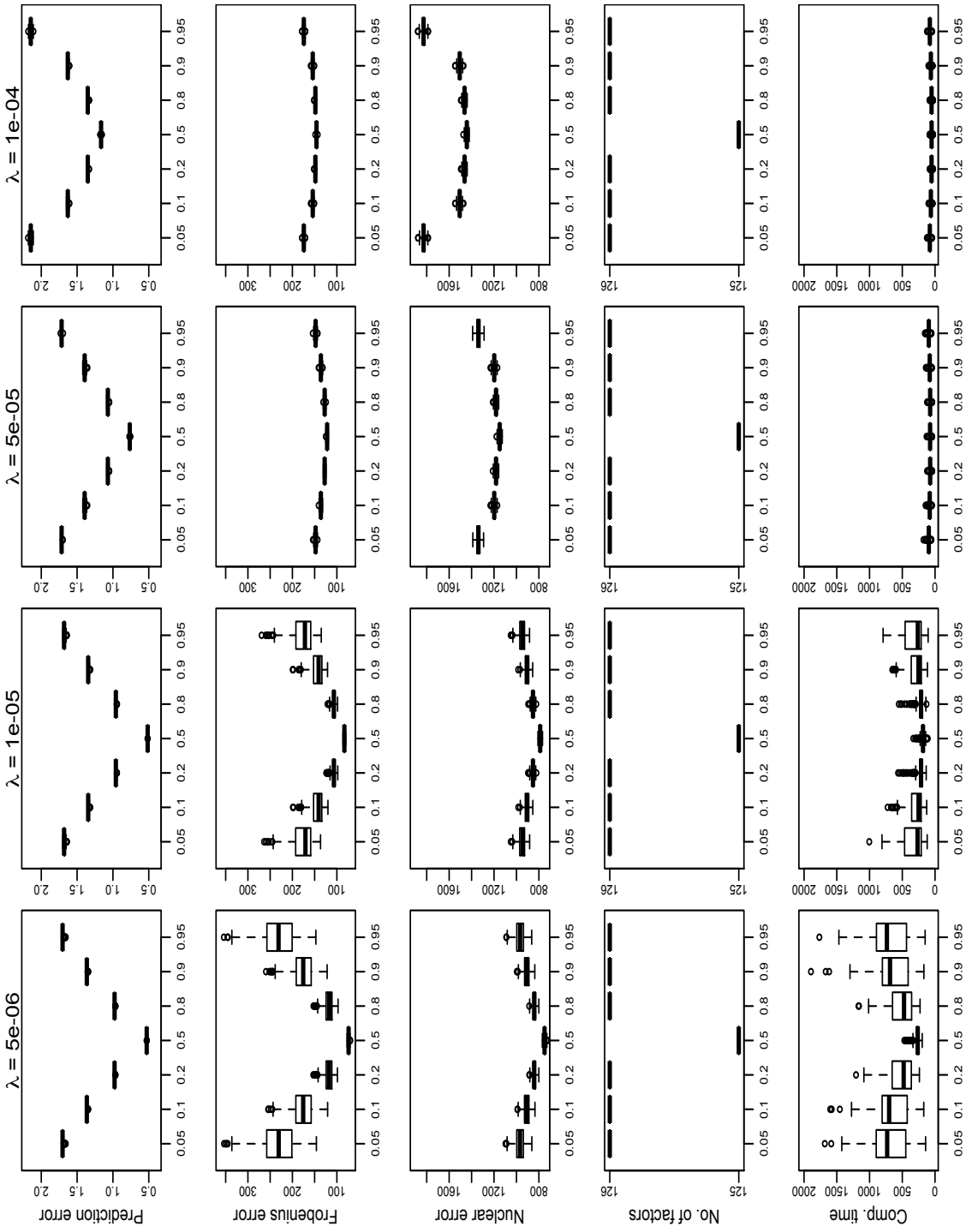


Figure 4.6.2: The symmetric Model LS. The horizontal axis is τ. The true number of factors is 125.

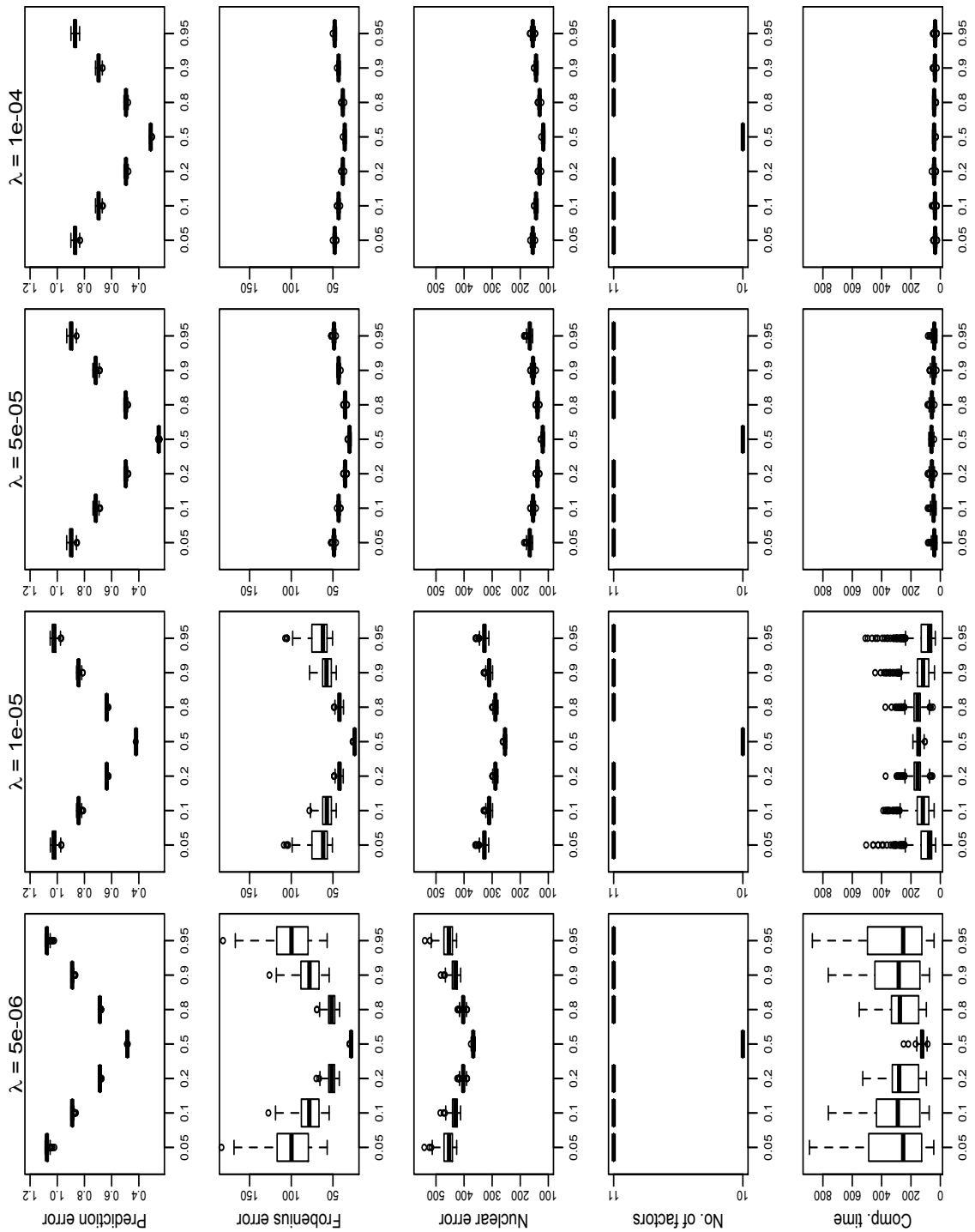


Figure 4.6.3: The symmetric Model MS. The horizontal axis is τ. The true number of factors is 10.

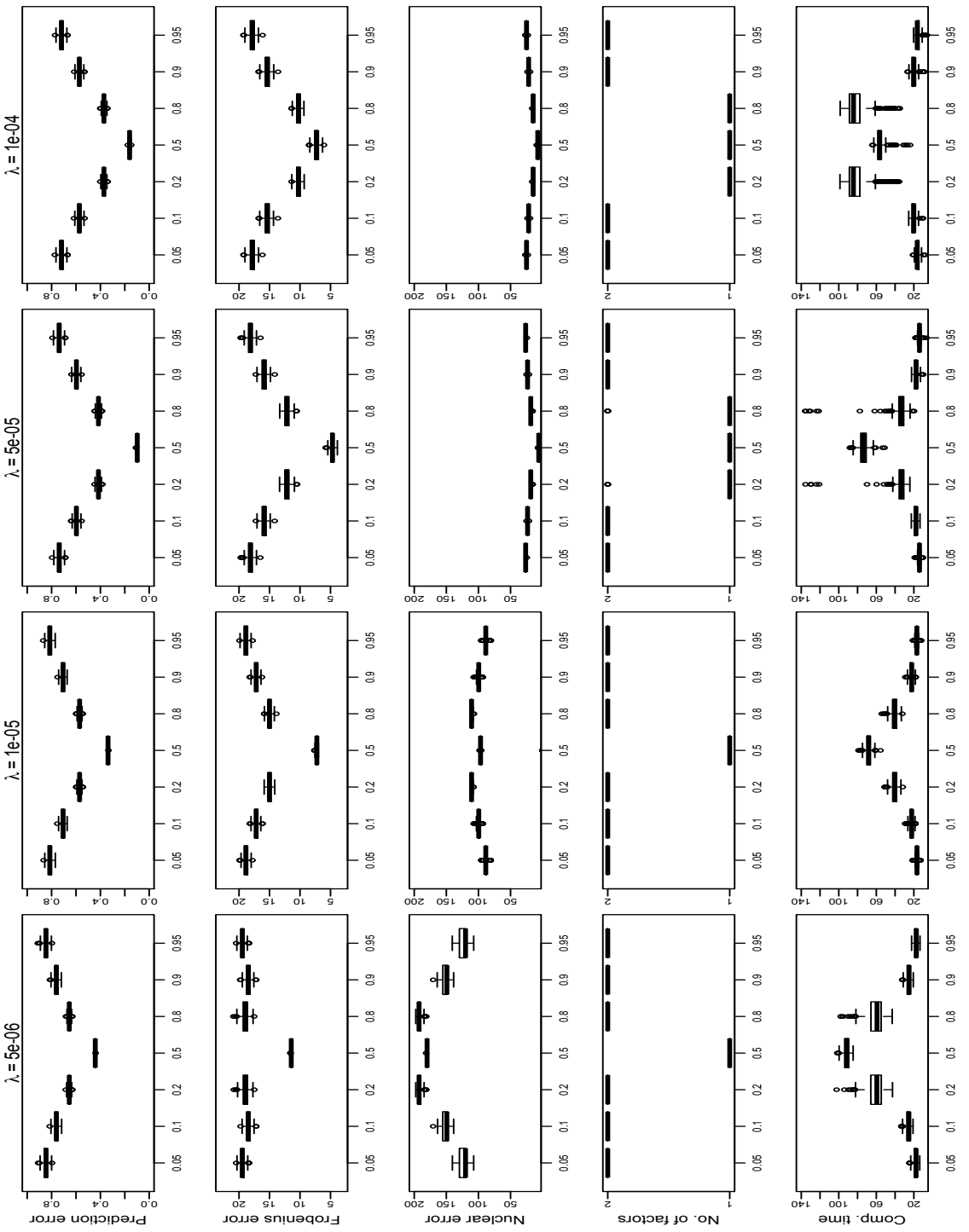


Figure 4.6.4: The symmetric Model ES. The horizontal axis is τ. The true number of factors is 1.

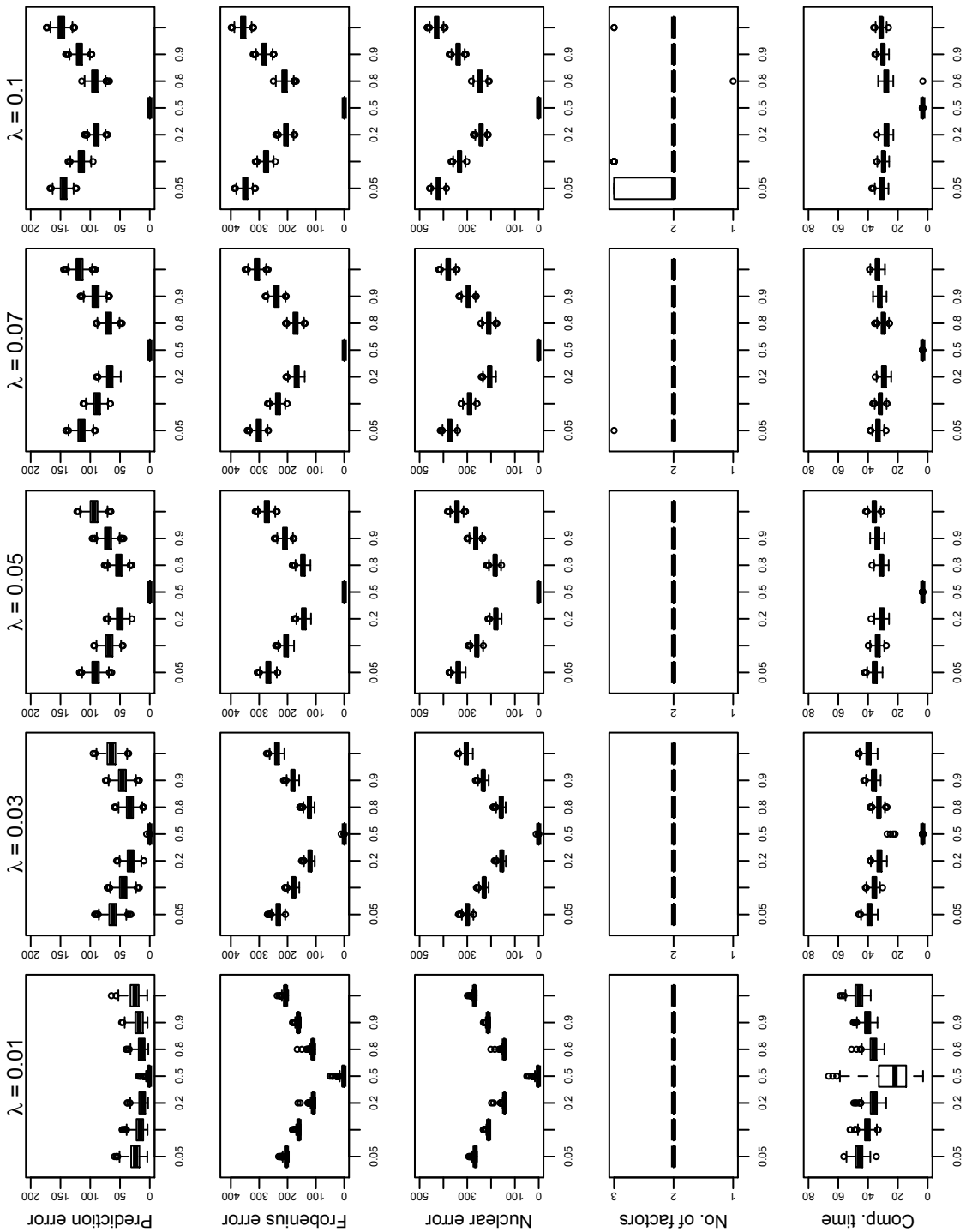


Figure 4.6.5: The asymmetric Model AES. The horizontal axis is τ. The true number of factors is 2 for $\tau < 0.5$ and 10 for $\tau > 0.5$. 0 for $\tau = 0.5$.

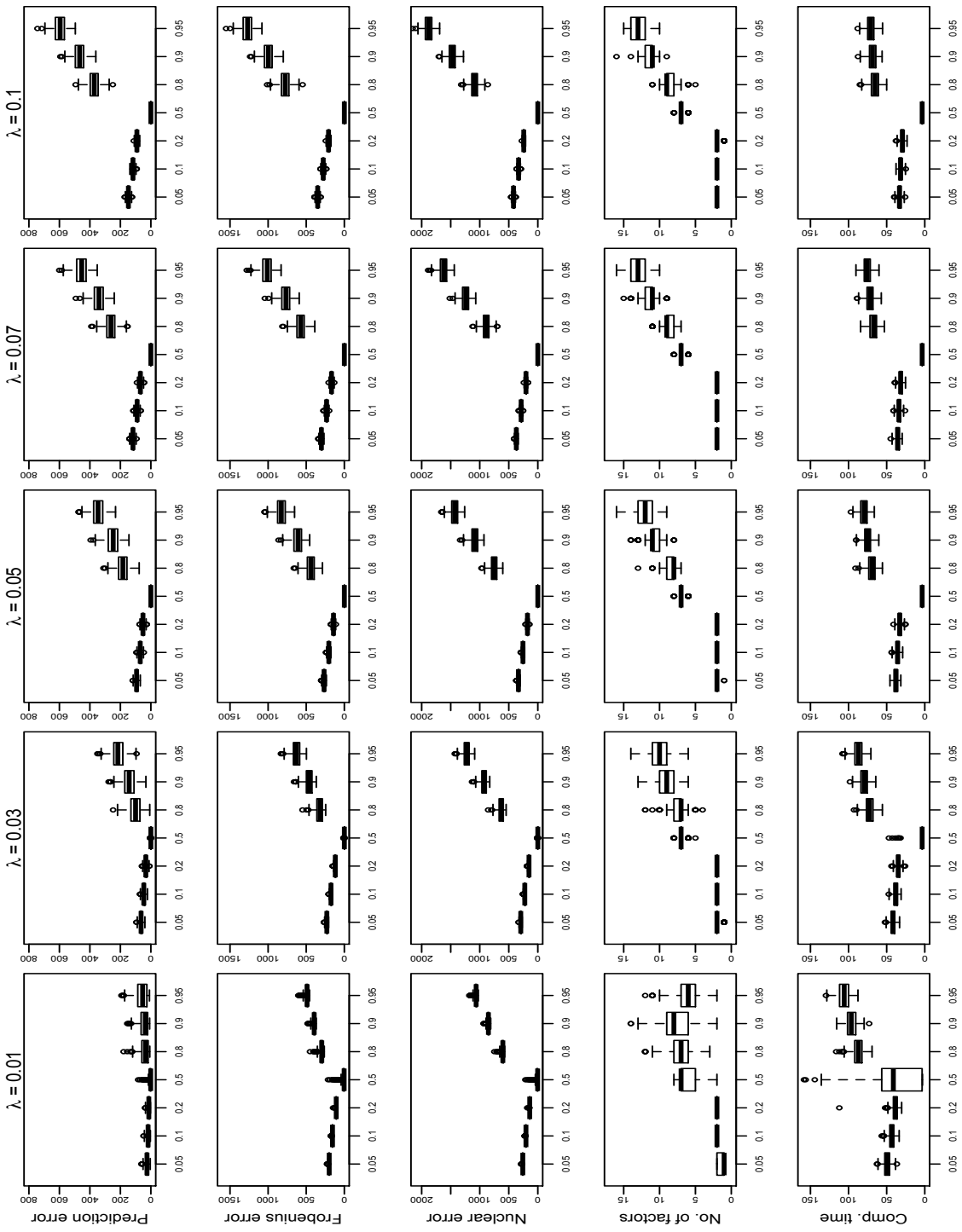


Figure 4.6.6: The asymmetric Model AMS. The horizontal axis is τ. The true number of factors is 2 for $\tau < 0.5$ and 10 for $\tau > 0.5$. 0 for $\tau = 0.5$.

4.7 Real data application: SAMCVaR model

In this section, we apply the regularized multiple quantile regression on financial data. In Section 4.7.1, we propose a modification of CAViaR model proposed by Engle and Manganelli (2004). Section 4.7.2 deals with the data selection and choice of the tuning parameter λ. Section 4.7.3 is devoted to the empirical findings.

4.7.1 Model

Since Engle and Manganelli (2004) proposed the conditional autoregressive value at risk (CAViaR) model around a decade ago, financial econometricians have applied it in many empirical studies and proposed many variations for it. This model is an autoregressive model in quantile, which does not account for the interdependence of asset returns. As the financial spillover effect has been widely understood as a major risk source, the quantification of spillover effect has been an important issue for financial econometricians.

White et al. (2008) introduce a multi-quantile modification of CAViaR (MQ-CAViaR), which allows a sequence of conditional quantile of asset returns to depend on each other. Combining with the robust estimation for skewness and kurtosis using quantiles of Kim and White (2004), they study the time varying patterns of higher moments of asset returns. In White et al. (2010), they consider the spillover effect in asset returns by the multivariate MQ-CAViaR (MVMQ-CAViaR) model, which combines the MQ-CAViaR models of a set of asset returns. Nonetheless, what they have actually done in their estimation is a single factor model. They estimated bivariate CAViaR for each asset with a single universal market index, for which they took the World Financials price index provided by Datastream.

In contrast to previous models, we consider a multivariate model which jointly incorporates all asset returns. Let $Y_{j,t}$ be the asset return for firm j, $j = 1, ..., m$, at time t, $t = 1, ..., T$. Let $q_{t,j}(\tau|\mathcal{F}_{t-1})$ be the conditional quantile at level τ for asset return j at time t on filtration $\mathcal{F}_{t-1}$. From the spirit of multivariate CAViaR, we consider the Sparse Asymmetric Multivariate Conditional Value-at-Risk model (SAMCVaR):

$$q_{t,j}(\tau|\mathcal{F}_{t-1}) = \sum_{k=1}^{m} \gamma_{1,j,k}(\tau)|Y_{t-1,k}| + \sum_{k=1}^{m} \gamma_{2,j,k}(\tau)Y_{t-1,k}^{-}, \tag{4.7.1}$$

where $Y^{-} = \max\{-Y, 0\}$, the coefficients $\mathbf{\Gamma}_{*j}(\tau) = (\boldsymbol{\gamma}_{1,j}(\tau)^{\top}, \boldsymbol{\gamma}_{2,j}(\tau)^{\top})^{\top}$ in which $\boldsymbol{\gamma}_{l,j}(\tau) = (\gamma_{l,j,1}(\tau), ..., \gamma_{l,j,m}(\tau))$ for $l = 1, 2$. The rank r of $\mathbf{\Gamma}$ satisfies $r \ll m$. Following the discussion in Section 4.2, we impose the condition that $\sum_{k=1}^{r} \psi_{j,k}^2 \leq 1$. Let

$$\boldsymbol{X}_{t-1} = (|Y_{t-1,1}|, ..., |Y_{t-1,m}|, Y_{t-1,1}^{-}, ..., Y_{t-1,m}^{-})^{\top} \in \mathbb{R}^{2m}. \tag{4.7.2}$$

We may therefore rewrite (4.7.1) as

$$q_{t,j}(\tau|\mathcal{F}_{t-1}) = q_{t,j}(\tau|\boldsymbol{X}_{t-1}) = \boldsymbol{X}_{t-1}^{\top}\mathbf{\Gamma}_{*j}(\tau).$$

If letting $\boldsymbol{q}_t(\tau|\boldsymbol{X}_{t-1}) = (q_{t,1}(\tau|\boldsymbol{X}_{t-1}), ..., q_{t,m}(\tau|\boldsymbol{X}_{t-1}))^\top$ be a vector of quantiles of all the firms in the sample, then $\boldsymbol{q}_t(\tau|\boldsymbol{X}_{t-1}) = \boldsymbol{\Gamma}^\top \boldsymbol{X}_{t-1}$, where $\boldsymbol{\Gamma} = [\boldsymbol{\Gamma}_{*1}, ..., \boldsymbol{\Gamma}_{*m}]$, and we have the multivariate quantile regression model (4.2.4)

This model is a multivariate variation of CAViaR, and we replace the autoregressive $q_{t-1,j}(\tau)$ in CAViaR model by a dispersion measure $|Y_{t-1,j}|$ for asset j in the information set at time $t-1$. The inclusion of the lag negative return $Y^-_{t-1,j}$, which also appears in the CAViaR model with "asymmetric slope", is based on the intuition that "one bad day makes the probability of the next somewhat greater" (Engle and Manganelli; 2004). Two major features of model (4.7.1) are that the quantile of each firm is *time-varying*, which is in the spirit of Engle and Manganelli (2004); moreover, model (4.7.1) accounts for the spillover effect on financial firm j from financial firm $l \neq j$.

We estimate $\boldsymbol{\Gamma}$ via the nuclear norm regularized multivariate quantile regression. We select $\tau = 1\%$ and 99%, in which $\tau = 1\%$ corresponds to the VaR of the asset returns, while $\tau = 99\%$ corresponds to the growth potential of the assets.

The *factorisation* described in Section 4.2 is applied to gain a deeper insight. We factorise the covariates into factors $f_{t,1}(\tau), ..., f_{t,r}(\tau)$ where $r \ll m$ by using the left singular vectors of $\boldsymbol{\Gamma}$. We investigate two aspects related to the factors. The first is how a firm $Y_{t-1,j}$ contributes to the factor; the second is how sensitive the conditional quantile of a firm is relative to the factor. We may study the contribution of firm j to the variation of the market by the coefficients associated to the two transformations $|Y_{tj}|, Y^-_{tj}$ in the factor f_k:

$$\text{Contribution from component } j \text{ to } f_k(\tau) : \frac{\partial f_k(\tau)}{\partial(|Y_j|, Y_j^-)} = (\varphi_{1,k,j}, \varphi_{2,k,j}). \quad (4.7.3)$$

Note that the contribution from component j to $f_r(\tau)$ does not vary over time. On the other hand, the sensitivity of relative to the variation of market can be described by

$$\text{Sensitivity of } j \text{ quantile to } f_k(\tau) : \frac{\partial q_j(\tau|\boldsymbol{X})}{\partial f_k(\tau)} = \psi_{j,k}. \quad (4.7.4)$$

With the singular value decomposition $\boldsymbol{\Gamma} = \mathbf{U}\mathbf{D}\mathbf{V}^\top$, the contribution of j firm to the factor f_k defined in (4.7.3) can be computed by the $j, j+m$ element in the $\mathbf{U}_{*k} \in \mathbb{R}^{2m}$ times σ_k, where σ_k is the kth singular value on the diagonal of $\mathbf{D}$. The quantity in (4.7.4) can be found by the kth component in $\mathbf{V}_{k*}$.

4.7.2 Data and tuning

We obtain a set of stock prices consists of $m = 230$ major global financial firms. The dataset can be downloaded from Simone Manganelli's website, which is used in White et al. (2010). Their data period is from Dec. 31, 1999 to Aug. 6, 2010. The regional and industrial characteristics can be found in Table 1 of White et al. (2010), which we include in Table 4.7.1 for completeness.

	Bank	Financial Service	Insurance	Total
EU	47	22	27	96
North America	25	17	28	70
Asia	47	14	3	64
Total	119	53	58	$m = 230$

Table 4.7.1: Summary of firm characteristics. There are three geographical categories: Europe, North America and Asia, and also three industrial categories: bank, financial service and insurance.

We use the data from August 31, 2007 to August 6, 2010. There are 766 closed price for each stock in the sample. We compute the daily log-return. This results in sample size $n = 765$. The dimension of the input variables $\boldsymbol{X}_t$ is $p = 2m = 460$, as we consider two transformations for each asset return, as in formula (4.7.2). Figure 4.7.1 shows the time series plots of the log-returns of the 230 financial institutions over this data period, and a plot of volatility index (VIX) kept by Chicago Board Options Exchange. The plot of asset returns suggests there are two large high volatility clusters before and after the beginning of the year 2009, which corresponds to the subprime mortgage crisis. Another phase of volatility increase is around mid 2010, which corresponds to the rising concern of the European debt crisis. The data show strong asymmetry, as the returns demonstrate high negative skewness. Though VIX mainly characterizes the volatility of the S&P500 constituents, it appears to be a good approximate for the global financial risk too.

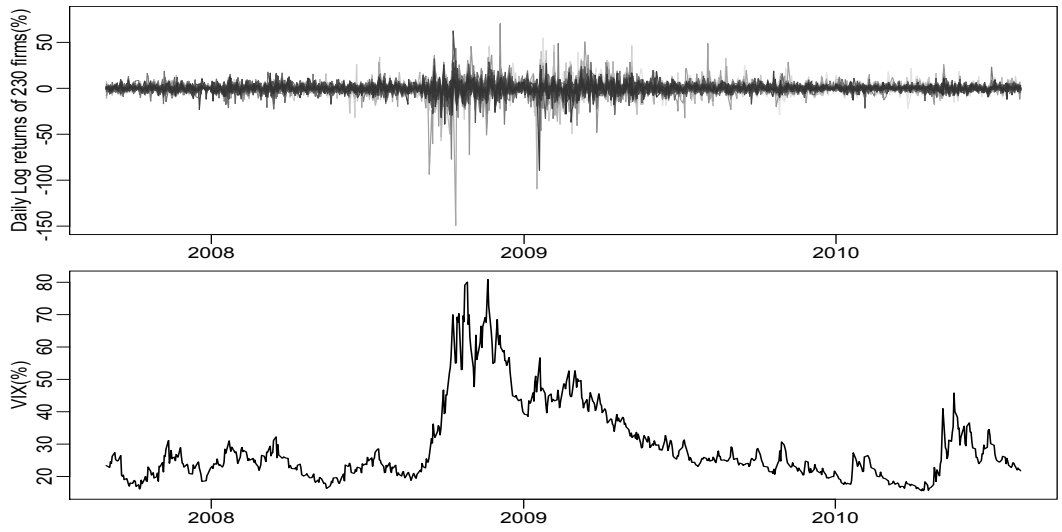


Figure 4.7.1: The upper figure shows the time series plots of the 230 global financial institutions with different grey level distributions and thicknesses. The lower figure shows the time series of VIX.

To select the tuning parameter λ, applying the procedure described in Section 4.5 gives $\lambda = 0.02467565$ for $\tau = 1\%$. By symmetry we also apply $\lambda = 0.02467565$

for $\tau = 99\%$.

4.7.3 Results

In this section we discuss the empirical findings from factorizing the multivariate quantile regression model (4.7.1) at level $\tau = 1\%$ and 99%. After the factorisation by SVD, the time series plot of the first two factors for the two set of quantile regression are reported in Figure 4.7.2. Both first factors $\boldsymbol{f}_1(0.01)$ and $\boldsymbol{f}_1(0.99)$ are volatile and moving away from 0 at the end of 2008 and in the first quarter of 2009, and mid 2010, which corresponds to the phases of volatility increase as indicated in Figure 4.7.1. Moreover, as can be seen from the figures, the two time series $\boldsymbol{f}_1(0.01)$ and $\boldsymbol{f}_1(0.99)$ are negatively correlated. The absolute scale of the two second factors $\boldsymbol{f}_2(0.01)$ and $\boldsymbol{f}_2(0.99)$ are much smaller than the first factors. A sharp peak appears in the plot of $\boldsymbol{f}_2(0.01)$ at the first quarter of 2009. The time series of $\boldsymbol{f}_2(0.99)$ is volatile before and after the beginning of the year 2009.

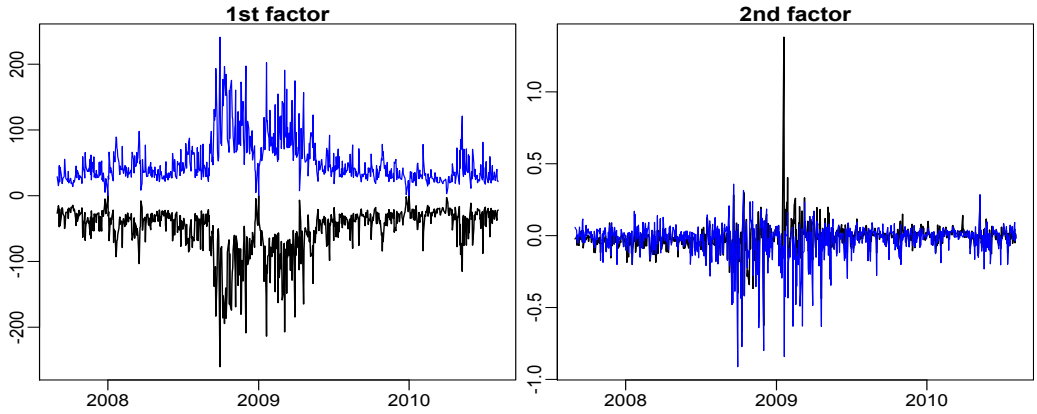


Figure 4.7.2: The time series plots for the first 2 factors. The black lines corresponds to 1% quantile factors and the blue lines corresponds to 99% quantile factors.

In what follows we discuss the risk contribution to the factors and the factor loadings of the firms in our sample. We begin with the first factor of 1% and 99% multivariate quantile regression. Figure 4.7.3 shows that the contribution to the first factors lie in the second quadrant, which suggests that all the covariates have negative impact to the first factor of 1% multivariate quantile regression, and positive impact to the first factor of 99% multivariate quantile regression. The black dots and black firm names represent the lag absolute log-returns, and they tend to lie around the diagonal line or even above it. This suggests that the absolute lag log-returns tend to contribute equally to both $\boldsymbol{f}_1(0.01)$ and $\boldsymbol{f}_1(0.99)$, which is consistent to the intuition that higher return is accompanied by higher risk. On the other hand, the lag negative part $Y^-_{t-1,j}$ marked in red are more located below the diagonal line, which suggests that the $Y^-_{t-1,j}$ contributes more to $\boldsymbol{f}_1(0.01)$ than to $\boldsymbol{f}_1(0.99)$. The well-known "leverage effect" postulated by Black (1976) suggests

the tendency that the volatility of an asset is negatively correlated to the the asset return. Furthermore, it is suggested that such effect is asymmetric: the association of losses with larger volatility changes than that of gains with lower volatility, as documented by Engle and Ng (1993). As volatility or variance is a symmetric measure of dispersion of distribution, it is incapable of revealing information of the potentially asymmetric contribution to such dispersion. However, Figure 4.7.3 uncovers the fact that the increasing dispersion of the distribution in asset return in response to the nonnegative loss $Y^-_{t-1,j}$ is largely due to the drop of lower quantile factor $\boldsymbol{f}_1(0.01)$ rather than the rise of upper quantile factor $\boldsymbol{f}_1(0.99)$. In particular, such increase in volatility does not create as much potential in gain as in loss.

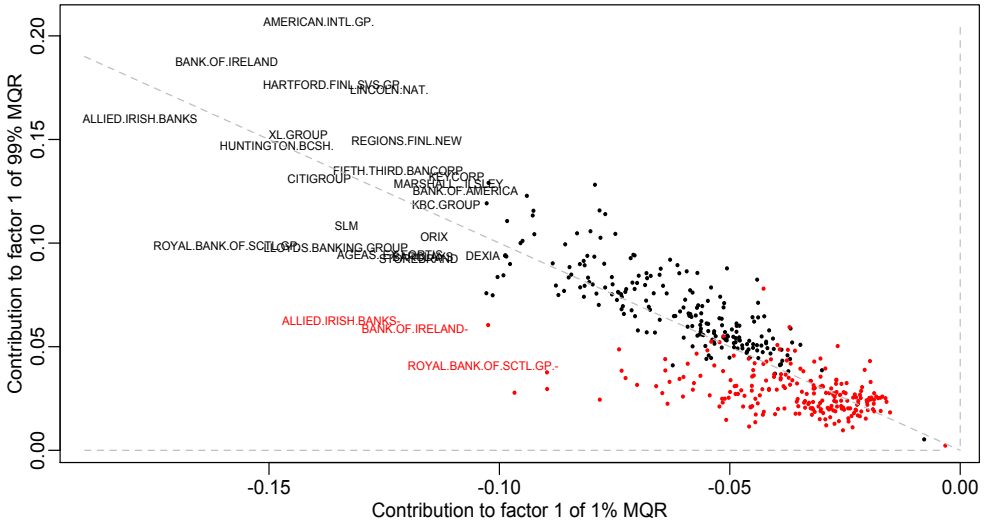


Figure 4.7.3: The magnitude of contribution to the first factor of 1% and 99% MQR from the 230+230 covariates. The firm name and the black dots denote the squared log return $Y^2_{t-1,j}$. Red dots and firm name with "–" denote the lag negative return $Y^-_{t-1,j}$.

Figure 4.7.4 illustrate the loadings to the first factor of of 1% and 99% multivariate quantile regression. The loadings are all positive, and lying on the 45 degree line, which suggests that the firm highly associated with the first factor of 1% MQR would also be highly associated with the first factor of 99% MQR. This implies that the trend of the τ-range of the returns is similar, but their magnitudes are different. Indeed, the firms lying on the far northeast are the firms with high market risk sensitivity, including Huntington Bancshares Inc., American International Group, Allied Irish Banks and more, whose time series patterns best resemble that of the first factors $\boldsymbol{f}_1(0.01)$ and $\boldsymbol{f}_1(0.99)$. The return time series of several risky firms are shown in Figure 4.7.10, in the sense that during financial crisis of 2008-2009, the range of their distribution is very disperse. Hence, their volatility is also very large.

Second factors $\boldsymbol{f}_2(0.99)$ and $\boldsymbol{f}_1(0.99)$ in Figure 4.7.5 suggest a different story from the first factors. The black dots are more located above the line corresponding

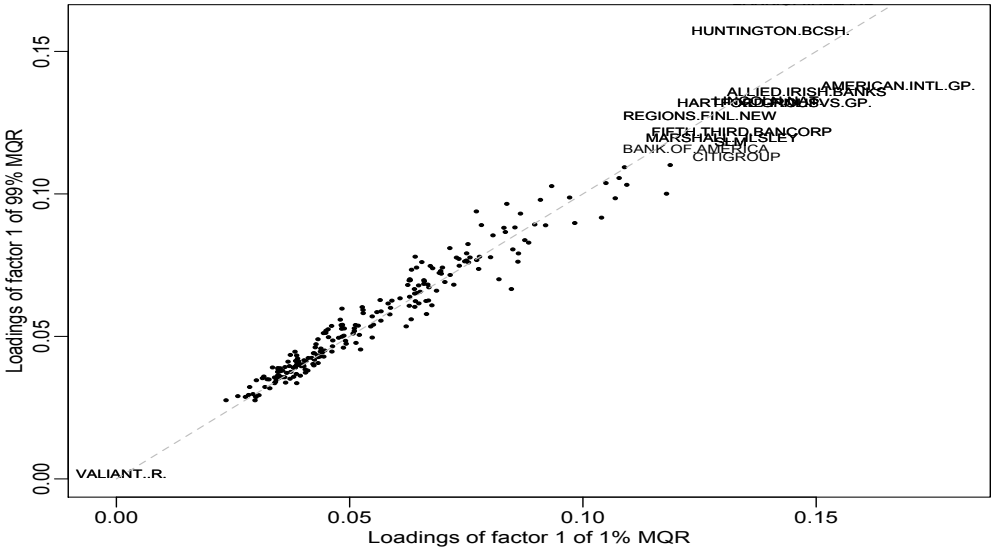


Figure 4.7.4: The factor loadings of 230 firms on the first factors $\boldsymbol{f}_1(0.01)$ and $\boldsymbol{f}_1(0.99)$.

to zero contribution to factor $\boldsymbol{f}_2(0.99)$, while the red dots tend to appear below the line. This suggests again that the negative part of lag log-returns have less to do with the factor associated with the upper quantile. Moreover, a few covariates are lying at far right and are separate from the other points, and these points are associated mainly with the Bank of Ireland, Allied Irish Banks and the Royal Bank of Scotland Group. It is these banks leading to the peak of $\boldsymbol{f}_2(0.99)$.

The factor loadings of firms on the second factor $\boldsymbol{f}_2(0.01)$ can be applied to distinguish the firms being influenced most by the sharp peak of $\boldsymbol{f}_2(0.01)$. In Figure 4.7.6, the loadings of asset returns on $\boldsymbol{f}_2(0.01)$ and $\boldsymbol{f}_2(0.99)$ are mainly distributed in the first and third quadrants. Those who suffer from the negative effect of the peak of $\boldsymbol{f}_2(0.99)$ are on the second and third quadrants. In particular, around the beginning of 2009, the second factors $\boldsymbol{f}_2(0.01)$ and $\boldsymbol{f}_2(0.99)$ impose on the firms in the second quadrant a "shift downward effect"; that is, the range of the return distribution of these firms shifts *downward* $\boldsymbol{f}_2(0.99)$ is mostly negative and $\boldsymbol{f}_2(0.01)$ has positive peak. However, the second factors impose on the first on the third quadrant a "expanding effect". Among the firms whose range expands the most are PNC Financial Services Group, Inc., State Street, Lloyds Banking Group PLC., and other large market participants.

The study of lower quantiles shows interesting result using the data during financial crisis. When the dimensionality of covariates is large, a reasonably good fitting to the quantile of response variables can be achieved with a few factors. Figure 4.7.7 illustrates the contribution from the covariates to the first and second factor of 1% MQR. Lag negative returns concentrates on the lower right of the figure and is below the horizontal line $y = 0$, and lag absolute returns spread around the horizontal line $y = 0$. The absolute and negative lag return of Allied Irish Bank, Bank of Royal Scotland Group and Bank of Ireland are more isolated and located in the

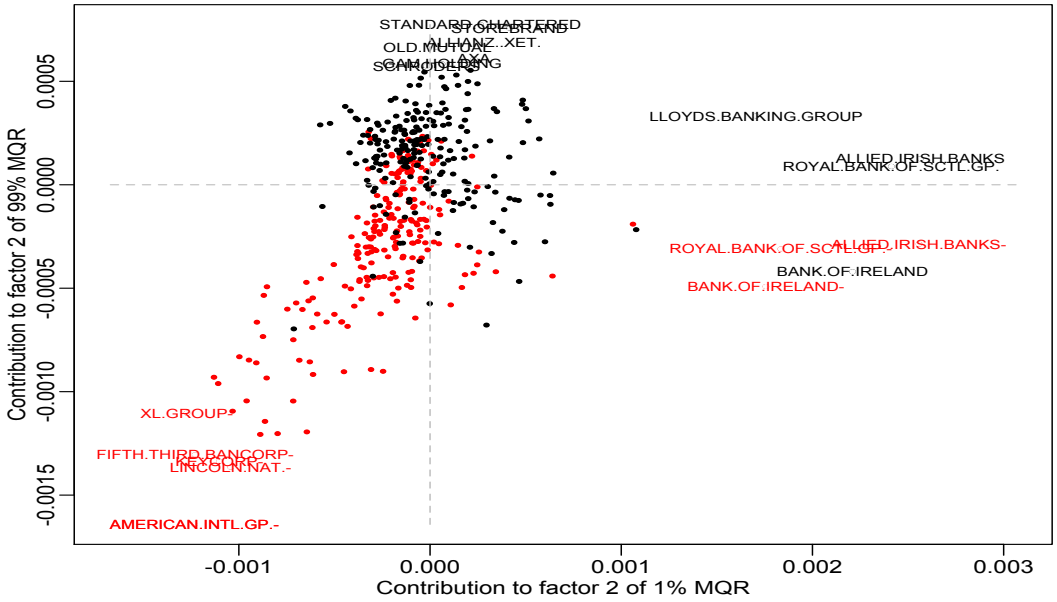


Figure 4.7.5: The magnitude of contribution to the second factor of 1% and 99% MQR from the 230+230 covariates. The firm name and the black dots denote the squared log return $Y^2_{t-1,j}$. Red dots and firm name with "–" denote the lag negative return $Y^-_{t-1,j}$.

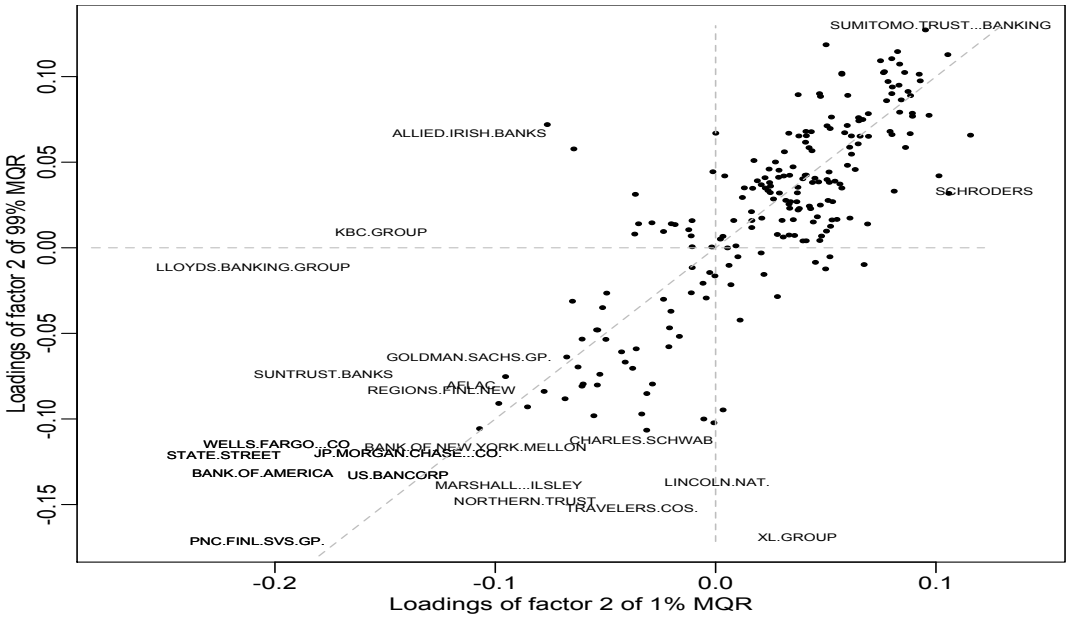


Figure 4.7.6: The factor loadings of 230 firms on the second factors $\boldsymbol{f}_2(0.01)$ and $\boldsymbol{f}_2(0.99)$.

top left corner, and are highly related to the first and second factor of 1% MQR. This suggests that they have high association with the global financial market.

Figure 4.7.8 shows the factor loadings of each firm on the first and second factors

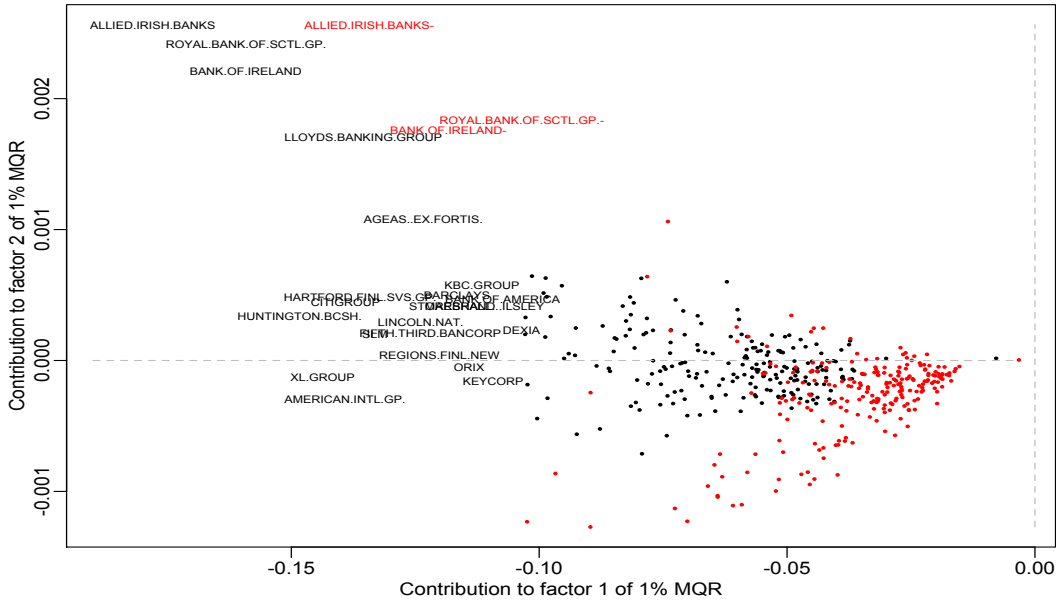


Figure 4.7.7: The magnitude of contribution to the *first* and *second* factor of 1% MQR from the 230+230 covariates. The firm name and the black dots denote the squared log return $Y^2_{t-1,j}$. Red dots and firm name with "–" denote the lag negative return $Y^-_{t-1,j}$.

of 1% MQR. The points are gathering on the top left with positive loadings on factor 2, and then spreading to the lower right like a fan. The pattern suggests that the firms positively associated with the first factor of 1% MQR tend to be negatively associated with the second factor of 1% MQR. This result is interesting because Figure 4.7.2 shows that the first factor of 1% MQR is generally negative and the second factor of 1% MQR has a positive peak. Hence, Figure 4.7.8 suggests that the firms lying on the lower right bear high market risk in our sample. Moreover, the shorter the distance between the two points on Figure 4.7.8, the larger their association is in their 1% quantile. That is, when one suffers losses, the other is likely to suffer losses by similar magnitude. For example, the distance between State Street and PNC Financial Services Group, Inc. is short, and their 1% quantile time series have similar behavior, which can also be seen from their time series plots in Figure 4.7.10.

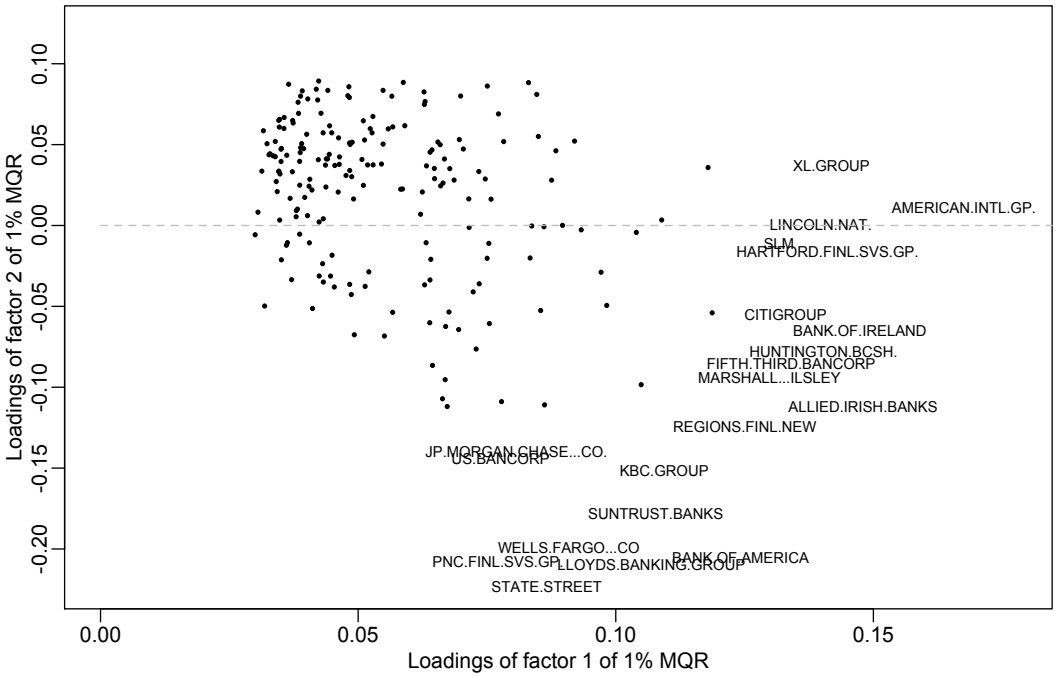


Figure 4.7.8: The factor loadings of 230 firms on the second factors $\boldsymbol{f}_1(0.01)$ and $\boldsymbol{f}_2(0.01)$ of 1% MQR.

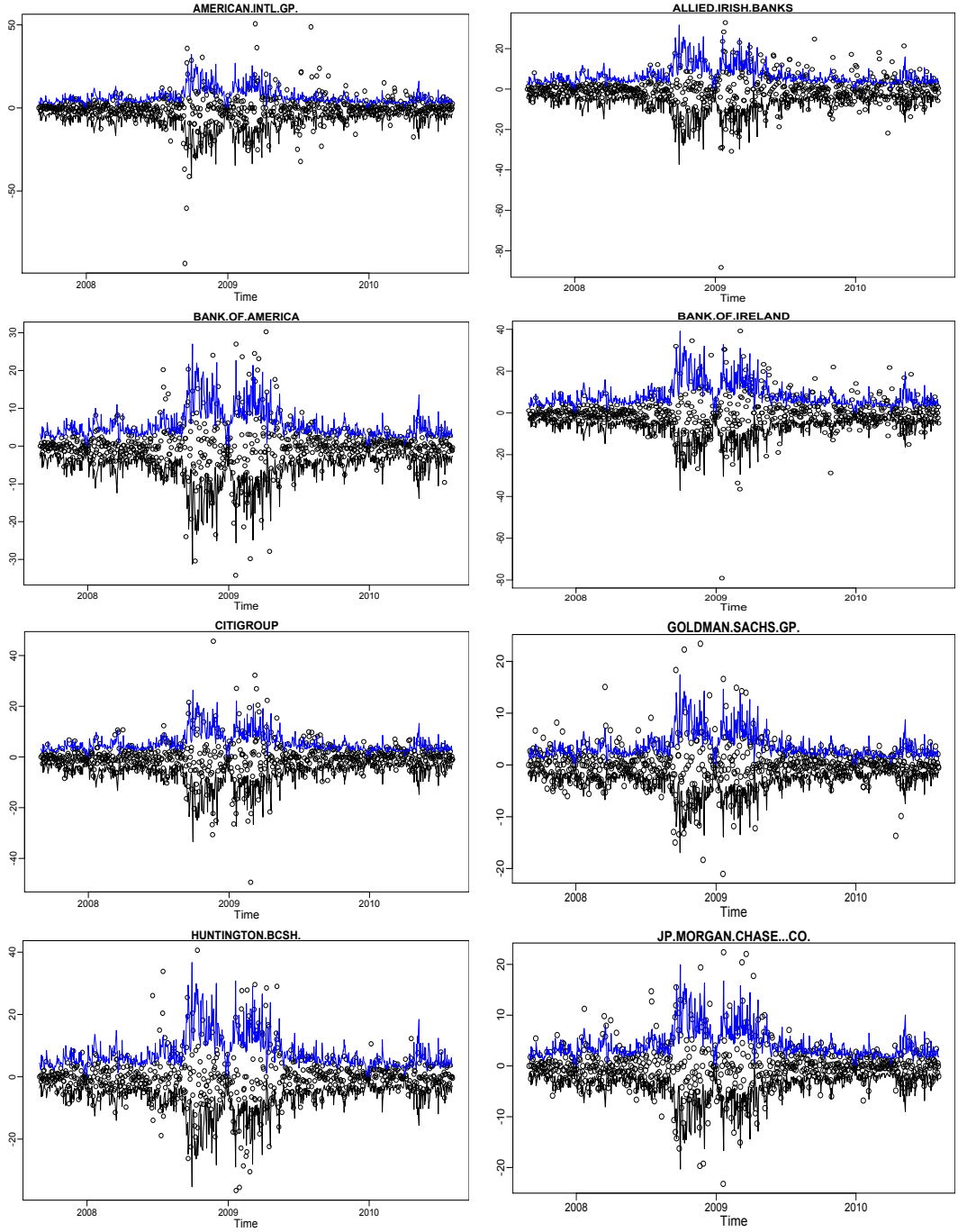


Figure 4.7.9: Plots of individual asset time series and their 1% and 99% fitted quantiles.

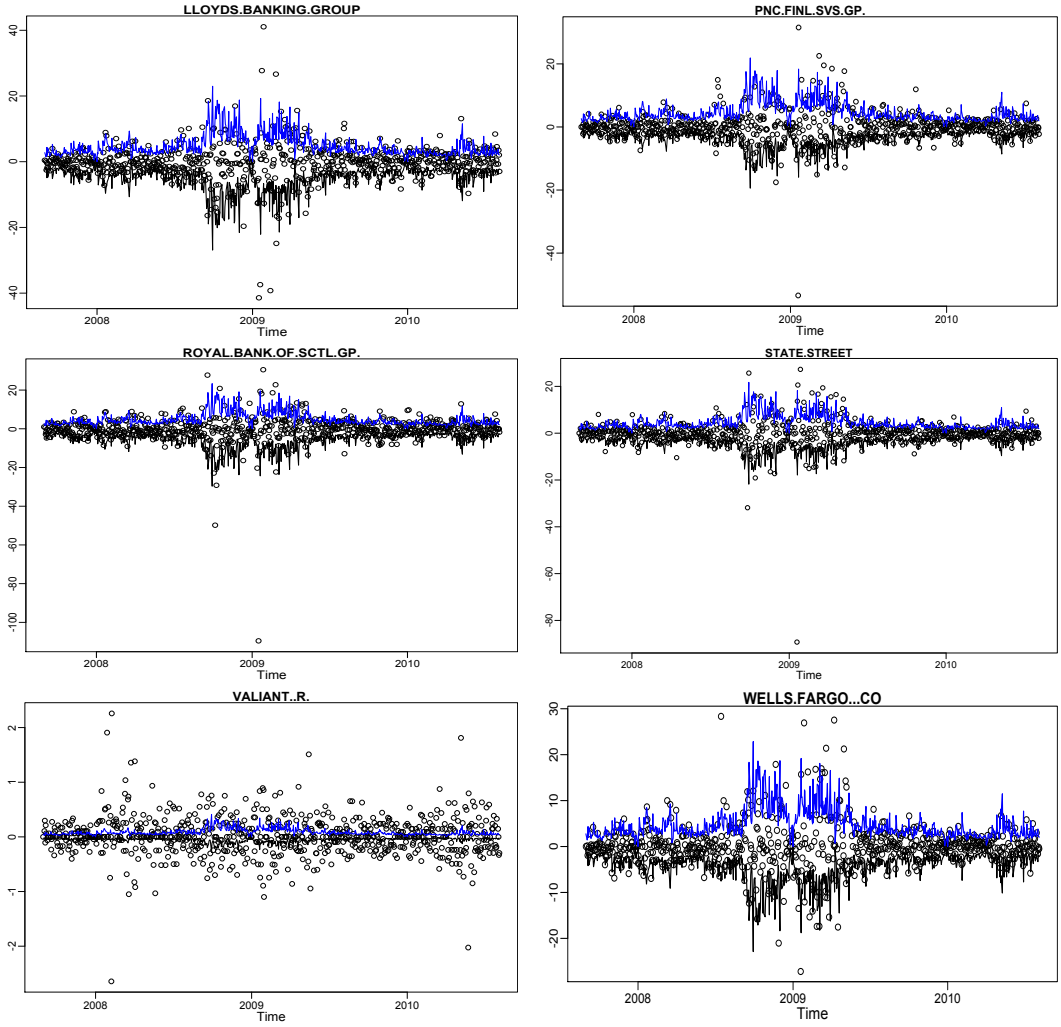


Figure 4.7.10: Plots of individual asset time series and their 1% and 99% fitted quantiles (continued).

4.8 Factor curve model

In this section, we extend the parametric linear multivariate quantile regression model to a nonparametric model, in which the unknown conditional curves are approximated by sieve spaces. Section 4.8.1 introduces the factorisable "quantile curve" and the factor curves. Section 4.8.2 deals with the estimation of the model. Section 4.8.3 applies the nonparametric factorisable quantile curves on the temperature data of 159 weather stations from China and classifies the primary patterns in Chinese temperature.

4.8.1 Model

For functional data, the concept of "quantile" is not as well understood as that for a usual univariate random variable. The functional data can be understood as the realizations of a *functional variable* (see, e.g. Ferraty and Vieu (2006)), which is a map $Y : \Omega \to \mathcal{C}$, where Ω is the sample space and $\mathcal{C}$ is the set of all continuous function on $\mathcal{T}$. Without loss of generality, $\mathcal{T}$ can be a bounded interval. As an example, the standard Brownian motion $W(\omega, t)$ is also a functional variable.

Definition 4.8.1 (Quantile Curves)**.** For $0 < \tau < 1$, the τ *quantile curve* $q_\tau(t)$ of functional variable Y is also a continuous function in t satisfying

$$\mathrm{P}\big(\{\omega : Y(\omega, t) \leq q_\tau(t),\ \forall t \in \mathcal{T}\}\big) = \tau.$$

For fixed $t \in \mathcal{T}$, it holds that $\mathrm{P}\big(\{\omega : Y(\omega, t) \leq q_\tau(t),\ \forall t \in \mathcal{T}\}\big) = \tau$. Taking standard Brownian motion $W(t)$ as an example, the τ quantile of $W(t)$ is $q_\tau(t) = \sqrt{t}\Phi^{-1}(\tau)$, where $\Phi(\cdot)$ is the cdf of standard normal distribution. When τ is close to 0 or 1, we call $q_\tau(t)$ a *tail event curve*.

Consider m functional variables $Y_1(t), ..., Y_m(t)$, denote their quantile curves $q_{\tau,j}(t)$. Suppose $q_{\tau,j}(t)$ lies in $\mathcal{F}$ which is the class of functions f defined on $[0, 1]$ whose sth derivative $f^{(s)}$ exists and satisfies a Lipschitz condition of order γ:

$$|f^{(s')}(t') - f^{(s')}(t)| \leq C|t' - t|^\theta, \quad \text{for } t', t \in [a, b],$$

for $s = s' + \theta > 0.5$. We assume that $s' \geq 1$ and $\theta > 0$ throughout the following discussion. Based on the construction of Schumaker (1981) and Stone (1985), each function $q_{\tau,j} \in \mathcal{F}$ can be approximated by an element $q_{n,\tau,j}(t) \in \mathcal{S}_n$ so that $\|q_{n,\tau,j} - q_{\tau,j}\|_\infty = \mathcal{O}(p_n^{-1})$ (see the discussion in p. 150 of Newey (1997)), where $\mathcal{S}_n$ is an expanding functional class with basis functions $\{b_l, 1 \leq l \leq p_n\}$. Denote $\boldsymbol{b}(t) = (b_1(t), ..., b_{p_n}(t))$, so that

$$q_{n,\tau,j}(t) = \boldsymbol{\Gamma}_{*j}^\top \boldsymbol{b}(t), \tag{4.8.1}$$

where $\boldsymbol{\Gamma}_{*j}$ is jth column of matrix $\boldsymbol{\Gamma}$.

The timing of measurement is $t_1, ..., t_n$ for all j. Denote $\mathbf{B} = (B_{il}) \in \mathbb{R}^{n \times p_n}$ with $B_{il} = b_l(t_i)$ and $\mathbf{Y} = (Y_{ij}) \in \mathbb{R}^{n \times m}$ with $Y_{ij} = Y_j(t_i)$. The matrix $\boldsymbol{\Gamma}$ can be viewed as the coefficient matrix in the multivariate quantile regression model

$$\boldsymbol{q}_{n,\tau}(t) = \mathbf{B}\boldsymbol{\Gamma}.$$

$\boldsymbol{q}_{n,\tau}(t) = (q_{n,\tau,1}(t), ..., q_{n,\tau,m}(t))$, and $\boldsymbol{\Gamma}$ can be estimated as in Section 4.3, but now the covariates are the values of the basis functions evaluated at $t_1, ..., t_n$.

Furthermore, model (4.8.1) is also *factorisable*. If the SVD of $\boldsymbol{\Gamma}$ is $\boldsymbol{\Gamma} = \mathbf{U}\mathbf{D}\mathbf{V}^\top$ and the number of singular values of $\boldsymbol{\Gamma}$ is r. Similarly to (4.2.5),

$$q_{n,\tau,j}(t) = \sum_{k=1}^{r} V_{j,k} f_k^\tau(t), \tag{4.8.2}$$

where $f_k^\tau(t) = \sigma_k \mathbf{U}_{*k}^\top \boldsymbol{b}(t)$ may be called *factor curves* with factor loadings $V_{j,k}$.

4.8.2 Estimation

Similar to Section 4.3, we minimize the following loss function:

$$(nm)^{-1} \sum_{i=1}^{n} \sum_{j=1}^{m} \rho_\tau\big(Y_{ij} - \mathbf{B}_{i*}^\top \boldsymbol{\Gamma}_{*j}\big) + \lambda \|\boldsymbol{\Gamma}\|_* \stackrel{\text{def}}{=} \widehat{Q}_{\tau,\boldsymbol{b}}(\boldsymbol{\Gamma}) + \lambda \|\boldsymbol{\Gamma}\|_*, \tag{4.8.3}$$

with $\rho_\tau(u) = |\mathbf{1}(u \leq 0) - \tau||u|$ with given $0 < \tau < 1$.

The empirical loss $\widehat{Q}_{\tau,\boldsymbol{b}}(\boldsymbol{\Gamma})$ is non-smooth. Apply the approach in Section 4.3, the smoothed version of $\widehat{Q}_{\tau,\boldsymbol{b}}(\boldsymbol{\Gamma})$ with a Lipschitz gradient is $\widehat{Q}_{\tau,\boldsymbol{b},\kappa}(\boldsymbol{\Gamma})$. Algorithm 2 can be directly applied by using $\widehat{Q}_{\tau,\boldsymbol{b},\kappa}(\boldsymbol{\Gamma})$. The convergence analysis is similar to Theorem 4.3.3.

Algorithm 2: Smoothing fast iterative shrinkage-thresholding algorithm (SFISTA)

Input: $\mathbf{Y}$, $\mathbf{B}$, λ, $\kappa = \frac{\epsilon}{2mn}$, $M = \frac{1}{\kappa m^2 n^2}\|\mathbf{B}\|^2$;
Initialization: $\boldsymbol{\Gamma}_0 = 0$, $\boldsymbol{\Omega}_1 = 0$, step size $\delta_1 = 1$;
for $t = 1, 2, ..., T$ **do**
  $\boldsymbol{\Gamma}_t = S_{\lambda/M}\left(\boldsymbol{\Omega}_t - \frac{1}{M}\nabla \widehat{Q}_{\tau,\boldsymbol{b},\kappa}(\boldsymbol{\Omega}_t)\right)$;
  $\delta_{t+1} = \frac{1+\sqrt{1+4\delta_t^2}}{2}$;
  $\boldsymbol{\Omega}_{t+1} = \boldsymbol{\Gamma}_t + \frac{\delta_t - 1}{\delta_{t+1}}(\boldsymbol{\Gamma}_t - \boldsymbol{\Gamma}_{t-1})$;
end
Output $\widehat{\boldsymbol{\Gamma}} = \boldsymbol{\Gamma}_T$

For the choice of the number of spline basis p_n, from bias and variance decomposition of spline estimator (Huang; 2003), under the fact that the functions to be estimated in our case are univariate, the convergence rate of the estimator is $\mathcal{O}_P(p_n^{-s} + \sqrt{p_n/n})$. The order of p_n minimizes the convergence rate is $n^{1/(2s+1)}$.

4.8.3 Application: Chinese temperature data

In this section we apply the nonparametric multivariate regression model to real data. The data we consider is the Chinese temperature data in the year 2008

from 159 weather stations around China, which is downloaded from the website of Research Data Center of CRC 649 of Humboldt-Universität zu Berlin. The dataset consists of one year time series of daily averaged temperature.

Before applying our method, we first fit a mean curve with smoothing spline which describes the mean temperature of China in the year 2008. In Figure 4.8.1, the bottom subfigure is the fitted trend curve, which shows seasonal pattern. The detrended temperature time series of 159 weather stations in the top figure of Figure 4.8.1 also demonstrate a seasonality pattern. The deviation to the mean temperature among these weather stations is larger in winter than in summer.

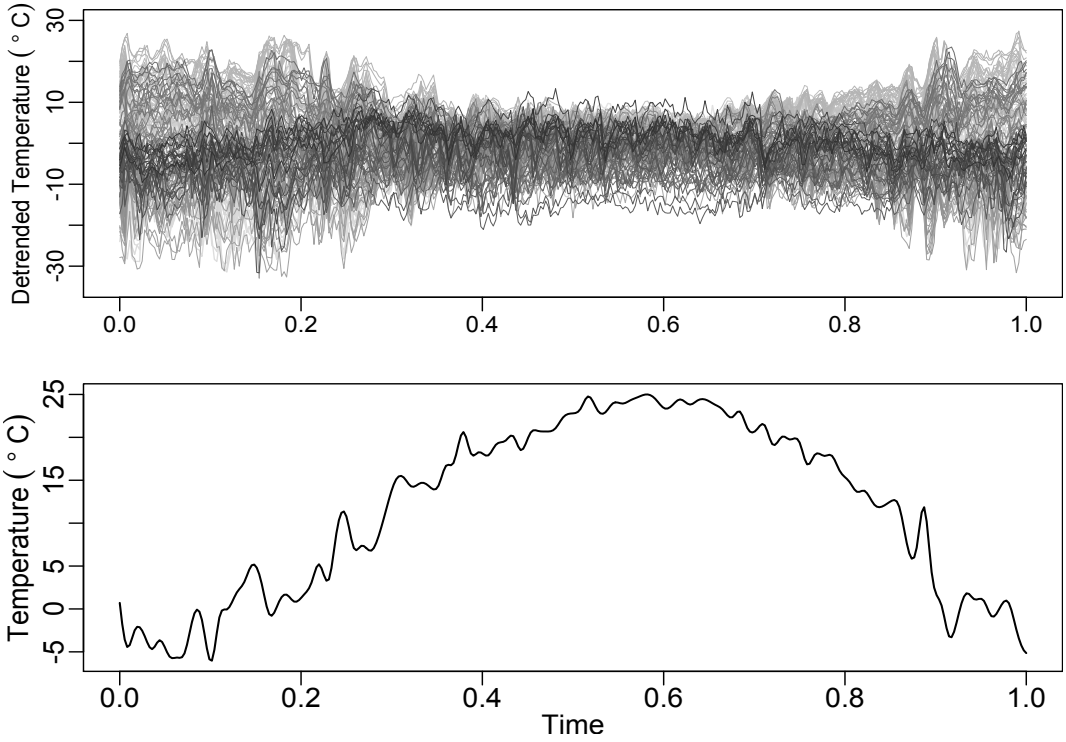


Figure 4.8.1: The temperature time series in excess to national mean of the 159 weather stations around China with different grey level distributions and thicknesses and the temperature trend curve.

We will apply the nonparametric multivariate quantile regression to further investigate the detrended temperature curves. The B-spline basis functions are used, and the number of basis function is $p = \lceil n^{0.4} \rceil = 11$. The timing of measurement is daily $t_1, ..., t_{365}$. The quantile levels are $\tau = 1\%$ and 99%. We choose the tuning parameter λ by applying the procedure of simulating (4.5.1) and compute λ by (4.5.2), the estimated value is $\lambda = 0.000156$.

Figure 4.8.2 presents the first four factors. The first factor of 1% and 99% quantile regression enclose a region which is wide in both ends and narrow in the middle. This matches our observation for Figure 4.8.1 that the deviation in temperature among weather stations tends to be higher in winter but lower in summer.

Moreover, the two first factors captures two types of seasonalities. The reverse V or U shape of the first factor of 99% multivariate quantile regression represents a "seasonality at high temperature", while the V or U shape of the first factor of 1% represents a "seasonality at low temperature". Note that we did not assume or impose any shape for the factors ex-ante. The shape of the factors are estimated by our algorithm.

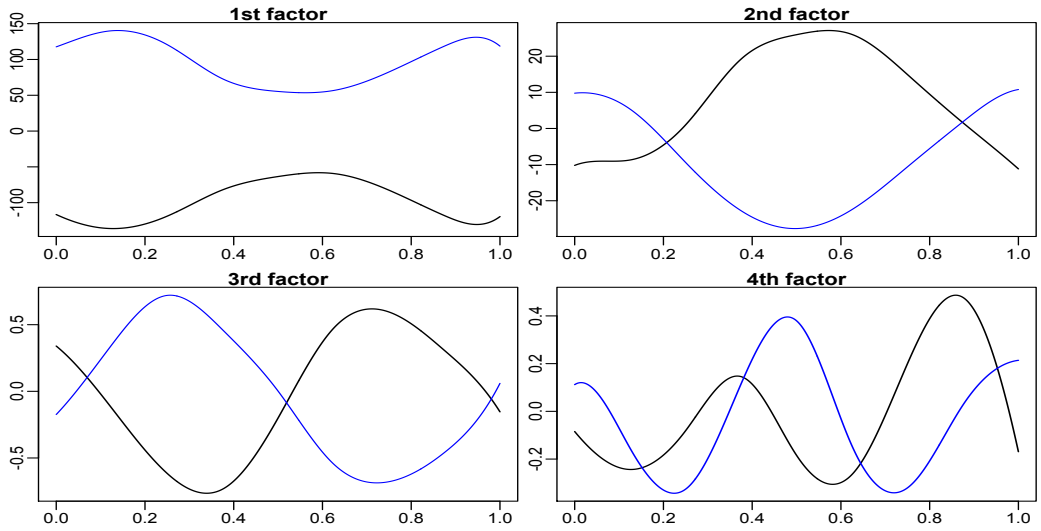


Figure 4.8.2: The time series plots for the first 4 factors. The black lines corresponds to 1% quantile factors and the blue lines corresponds to 99% quantile factors.

The factor loadings of the first factor for 1% and 99% quantile regression demonstrate a nearly "L" shape, as shown in Figure 4.8.3. This suggests that the weather stations nonnegatively associated with the first factor of 1% multivariate quantile regression have almost no association with the first factor of 99% multivariate quantile regression. Such dichotomy pattern allows for classifying the weather stations into groups.

In Figure 4.8.3, the temperature curve of Tulihe has the highest factor loading in the first factor of 1% multivariate quantile regression, while the temperature curve of Dongfang has the highest factor loading in the first factor of 99% multivariate quantile regression. Thus, Tulihe is classified as showing strong "seasonality at low temperature" and Dongfang shows strong "seasonality at high temperature". Notice that the factor loading to the first factor of 99% multivariate quantile regression is close to zero or slightly negative for Tulihe, and the factor loading to the first factor of 1% multivariate quantile regression is close to 0 for Dongfang. Another weather station marked in the figure is located in Yushu, which has small positive loadings to the first factor of both 1% and 99% multivariate quantile regression, and is hard to be classified to any of the two seasonality patterns.

Figure 4.8.4 shows the temperature plot, 1% and 99% quantile curves, and the location of the three weather stations marked in Figure 4.8.3. Tulihe is located in far

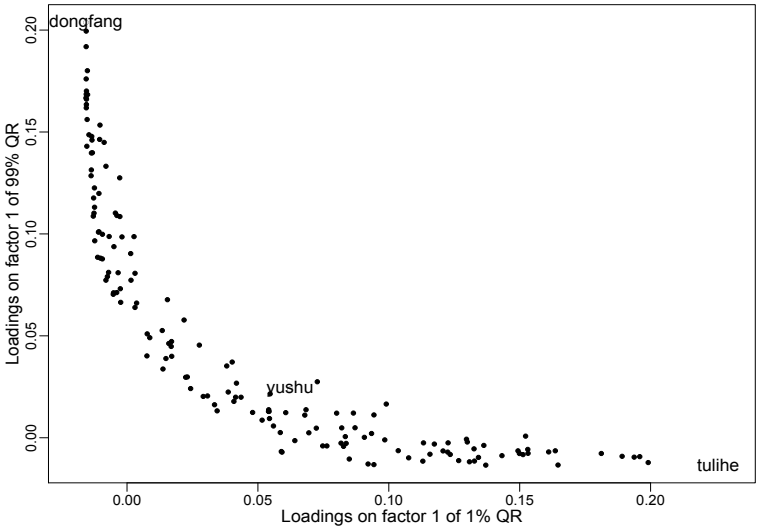


Figure 4.8.3: The plot of weather stations based on their factor loadings to 1% and 99% multivariate quantile regression. Each point denotes a weather station somewhere in China.

northeastern Inner Mongolia, China, which is well-known for its chilliness in winter and large temperature difference between summer and winter. The estimated 99% factors are mainly influenced by the temperature curves from warmer areas. Therefore, the reverse V-shaped yearly temperature curve of Tulihe cannot be estimated by the 99% factors, and the estimated curve is flat. Dongfang, however, is located in tropics, and in winter at warmest the temperature is 25 degrees celsius higher than the national average. The estimated 1% factors are incapable of forming the V-shaped temperature curve of Dongfang, so its 1% quantile curve is flat. Yuchu is located in central west China and belongs to highland climate. The average altitude in the region of Yuchu is over 4000 meters. It has high temperature variation within a day, and is generally slightly cooler in summer and warmer in winter than the national average. The seasonality for Yuchu is not significant.

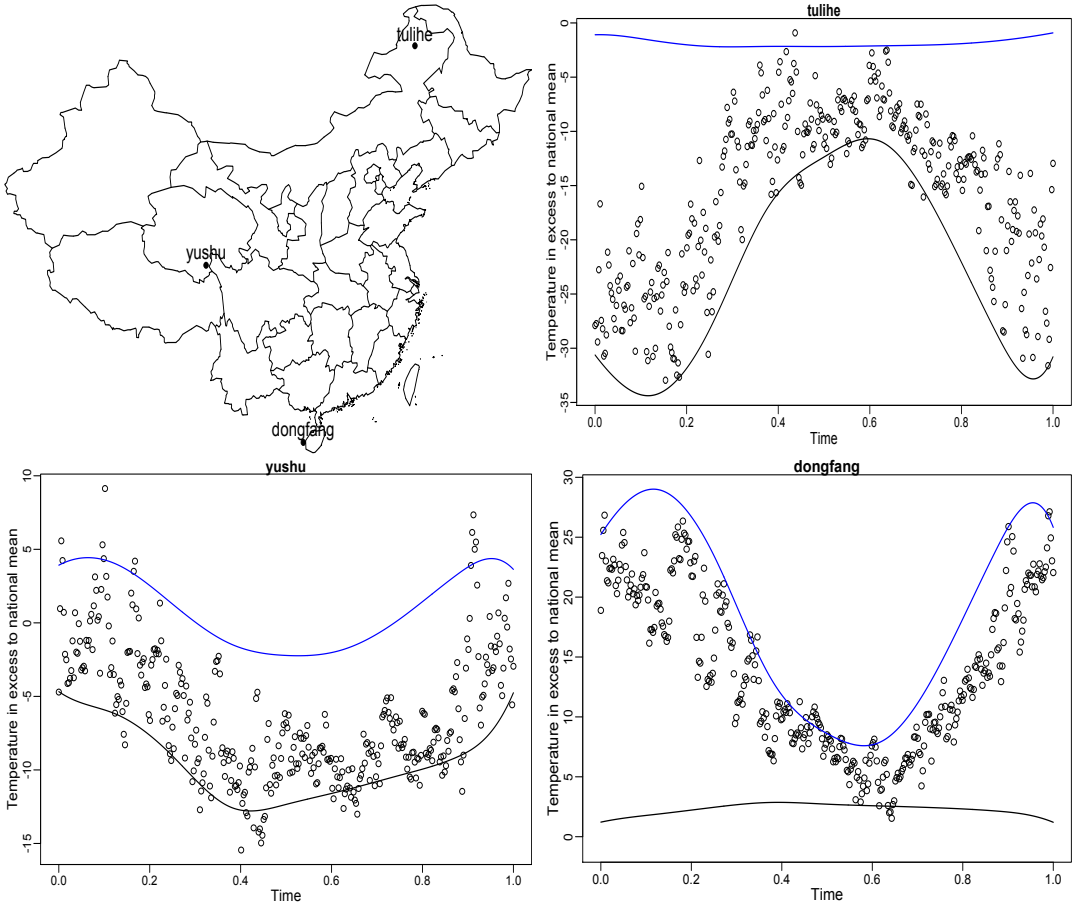


Figure 4.8.4: Plots of temperature observations, 1%, and 99% temperature quantile curves of the three weather stations in the year 2008. The location of the weather stations are marked in the upper left map of China.

Bibliography

Acharya, V. V., Pedersen, L. H., Philippon, T. and Richardson, M. (2010). Measuring systemic risk, *Working Paper 10-02*, Federal Reserve Bank of Cleveland.

Adams, Z., Füss, R. and Gropp, R. (2010). Modeling spillover effects among financial institutions: A State-Dependent Sensitivity Value-at-Risk (SDSVaR) approach, *Research Paper 10-12*, European Business School.

Adrian, T. and Brunnermeier, M. K. (2011). CoVaR, *Staff Reports 348*, Federal Reserve Bank of New York.

Anderson, T. W. (1951). Estimating linear restrictions on regression coefficients for multivariate normal distributions, *Annals of Mathematical Statistics* **22**: 327–351.

Bae, K.-H., Karolyi, G. A. and Stulz, R. M. (2003). A new approach to measuring financial contagion, *The Review of Financial Studies* **16**(3): 717–763.

Basel accords (2011). Basel III: A global regulatory framework for more resilient banks and banking systems, *Technical report*, Bank of International Settlements.

Beck, A. and Teboulle, M. (2009). A fast iterative shrinkage-thresholding algorithm for linear inverse problems, *SIAM Journal on Imaging Sciences* **2**(1): 183–202.

Belloni, A. and Chernozhukov, V. (2011). ℓ_1-penalized quantile regression in high-dimensional sparse models, *The Annals of Statistics* **39**(1): 82–130.

Belloni, A., Chernozhukov, V. and Wang, L. (2011). Square-root lasso: pivotal recovery of sparse signals via conic programming, *Biometrika* **98**(4): 791–806.

Berkowitz, J., Christoffersen, P. and Pelletier, D. (2011). Evaluating value-at-risk models with desk-level data, *Management Science* **57**(12): 2213–2227.

Bickel, P. J., Ritov, Y. and Tsybakov, A. B. (2009). Simultaneous analysis of Lasso and Dantzig selector, *The Annals of Statistics* **37**(4): 1705–1732.

Bickel, P. J. and Rosenblatt, M. (1973). On some global measures of the deviations of density function estimates, *The Annals of Statistics* **1**(6): 1071–1095.

Bickel, P. J. and Wichura, M. J. (1971). Convergence criteria for multivariate stochastic processes and some applications, *The Annals of Mathematical Statistics* **42**(5): 1656–1670.

Bissantz, N., Dümbgen, L., Holzmann, H. and Munk, A. (2007). Nonparametric confidence bands in deconvolution density estimation, *Journal of the Royal Statistical Society: Series B* **69**(3): 483–506.

Black, F. (1976). Studies of stock market volatility changes, *Proceedings of the American Statistical Association*, Business and Economic Statistics, pp. 177–181.

Brownlees, C. T. and Engle, R. (2010). Volatility, correlation and tails for systemic risk measurement, *Working paper*, NYU Stern School of Business.

Brunnermeier, M. and Pedersen, L. H. (2008). Market liquidity and funding liquidity, *Review of Financial Studies* **22**: 2201–2238.

Bunea, F., She, Y. and Wegkamp, M. H. (2011). Optimal selection of reduced rank estimators of high-dimensional matrices, *The Annals of Statistics* **39**(2): 1282–1309.

Cai, J.-F., Candès, E. J. and Shen, Z. (2010). A singular value thresholding algorithm for matrix completion, *SIAM Journal on Optimization* **20**(4): 1956–1982.

Cai, Z. and Wang, X. (2008). Nonparametric estimation of conditional VaR and expected shortfall, *Journal of Econometrics* **147**: 120–130.

Carroll, R. and Härdle, W. (1989). Symmetrized nearest neighbor regression estimates, *Statistics and Probability Letters* **7**: 315–318.

Chakraborty, B. (2003). On multivariate quantile regression, *Journal of Statistical Planning and Inference* **110**: 109–132.

Chaudhuri, P. (1996). On a geometric notion of quantiles for multivariate data, *Journal of American Statistical Association* **91**(434): 862–872.

Chen, X., Lin, Q., Kim, S., Carbonell, J. G. and Xing, E. P. (2012). Smoothing proximal gradient method for general structured sparse regression, *The Annals of Applied Statistics* **6**(2): 719–752.

Chernozhukov, V. and Umantsev, L. (2001). Conditional value-at-risk: Aspects of modeling and estimation, *Empirical Economics* **26**: 271–292.

Claeskens, G. and Van Keilegom, I. (2003). Bootstrap confidence bands for regression curves and their derivatives, *The Annals of Statistics* **31**(6): 1852–1884.

Dedecker, J., Merlevéde, F. and Rio, E. (2014). Strong approximation of the empirical distribution function for absolutely regular sequences in $\mathbb{R}^d$, *Electronic Journal of Probability* **19**(9): 1–56.

Dehejia, R. H. and Wahba, S. (1999). Causal effects in nonexperimental studies: Reevaluating the evaluation of training programs, *Journal of the American Statistical Association* **94**(448): 1053–1062.

Delgado, M. A. and Escanciano, J. C. (2013). Conditional stochastic dominance testing, *Journal of Business & Economic Statistics* **31**(1): 16–28.

Doksum, K. (1974). Empirical probability plots and statistical inference for nonlinear models in the two-sample case, *The Annals of Statistics* **2**(2): 267–277.

Engle, R. F. and Ng, V. K. (1993). Measuring and testing the impact of news on volatility, *The Journal of Finance* **48**(5): pp. 1749–1778.
URL: *http://www.jstor.org/stable/2329066*

Engle, R. and Manganelli, S. (2004). CAViaR: Conditional autoregressive value at risk by regression quantiles, *Journal of Business & Economic Statistics* **22**: 367–381.

Falk, M. (1999). A simple approach to the generation of uniformly distributed random variables with prescribed correlation, *Communications in Statistics - Simulation and Computation* **28**(3): 785–791.

Fan, J., Hu, T.-C. and Truong, Y. K. (1994). Robust nonparametric function rstimation, *Scandinavian Journal of Statistics* **21**: 433–446.

Fan, J., Xue, L. and Zou, H. (2013). Multi-task quantile regression under the transnormal model.

Fan, Y. and Liu, R. (2013). A direct approach to inference in nonparametric and semiparametric quantile regression models, Preprint.

Feng, X., He, X. and Hu, J. (2011). Wild bootstrap for quantile regression, *Biometrika* **98**(4): 995–999.

Ferraty, F. and Vieu, P. (2006). *Nonparametric Functional Data Analysis: Theory and Practice*, Springer.

Gibbons, M. and Ferson, W. (1985). Testing asset pricing models with changing expectations and an unobservable market portfolio, *Journal of Financial Economics* **14**: 217–236.

Giné, E. and Nickl, R. (2010). Confidence bands in density estimation, *The Annals of Statistics* **38**(2): 1122–1170.

Guerre, E. and Sabbah, C. (2012). Uniform bias study and Bahadur representation for local polynomial estimators of the conditional quantile function, *Econometric Theory* **28**(1): 87–129.

Guo, M. and Härdle, W. (2012). Simultaneous confidence bands for expectile functions, *AStA Advances in Statistical Analysis* **96**(4): 517–541.

Hall, P. (1979). On the rate of convergence of normal extremes, *Journal of Applied Probability* **16**(2): 433–439.

Hall, P. (1991). On convergence rates of suprema, *Probability Theory and Related Fields* **89**(4): 447–455.

Hall, P. (1992). Effect of bias estimation on coverage accuracy of bootstrap confidence intervals for a probability density, *The Annals of Statistics* **20**(2): 675–694.

Hall, P. and Horowitz, J. (2013). A simple bootstrap method for constructing nonparametric confidence bands for functions, *The Annals of Statistics* **41**(4): 1892–1921.

Hallin, M., Paindaveine, D. and Šiman, M. (2010). Multivariate quantiles and multiple-output regression quantiles: From L_1 optimization to halfspace depth, *The Annals of Statistics* **38**(2): 635–669.

Hansen, B. E. (2008). Uniform convergence rates for kernel estimation with dependent data, *Econometric Theory* **24**(3): 726–748.

Härdle, W. (1989). Asymptotic maximal deviation of M-smoothers, *Journal of Multivariate Analysis* **29**(2): 163–179.

Härdle, W., Liang, H. and Gao, J. (2000). *Partially Linear Models*, Physica-Verlag, Heidelberg.

Härdle, W., Müller, M., Sperlich, S. and Werwatz, A. (2004). *Nonparametric and Semiparametric Models*, Springer-Verlag, Berlin.

Härdle, W. and Song, S. (2010). Confidence bands in quantile regression, *Econometric Theory* **26**: 1180–1200.

Härdle, W., Spokoiny, V. and Wang, W. (2013). Local quantile regression, *Journal of Statistical Planning and Inference* **143**(7): 1109–1129.

Hautsch, N., Schaumburg, J. and Schienle, M. (2014). Financial network systemic risk contributions, *Review of Finance* pp. 1–54.

Hazan, E. (2008). Sparse approximate solutions to semidefinite programs, *LATIN 2008: Theoretical Informatics*.

Huang, J. Z. (2003). Local asymptotics for polynomial spline regression, *Annals of Statistics* **31**(5): 1600–1635.

Huang, X., Zhou, H. and Zhu, H. (2011). Systemic risk contributions, *Staff working papers 2011-08*, The Federal Reserve Board.

Izenman, A. J. (1975). Reduced-rank regression for the multivariate linear model, *Journal of Multivariate Analysis* **5**: 248–264.

Jaggi, M. and Sulovský, M. (2010). A simple algorithm for nuclear norm regularized problems, *Proceedings of the 27th International Conference on Machine Learning*.

Ji, S. and Ye, J. (2009). An accelerated gradient method for trace norm minimization, *Proceedings of the 26th International Conference on Machine Learning*.

Johnston, G. J. (1982). Probabilities of maximal deviations for nonparametric regression function estimates, *Journal of Multivariate Analysis* **12**(3): 402–414.

Jones, M. C. (1994). Expectiles and M-quantiles are quantiles, *Statistics & Probability Letters* **20**(2): 149–153.

Kim, T.-H. and White, H. (2004). On more robust estimation of skewness and kurtosis, *Finance Research Letters* **1**: 56–73.

Kiwitt, S. and Neumeyer, N. (2012). Estimating the conditional error distribution in non-parametric regression, *Scandinavian Journal of Statistics* **39**(2): 259–281.

Koenker, R. (2005). *Quantile Regression*, Econometric Society Monographs, Cambridge University Press, New York.

Koenker, R. and Bassett, G. S. (1978). Regression quantiles, *Econometrica* **46**(1): 33–50.

Koenker, R. and Portnoy, S. (1990). M estimation of multivariate regressions, *Journal of American Statistical Association* **85**(412): 1060–1068.

Koltchinskii, V. (2013). Sharp oracle inequalities in low rank estimation, *in* B. Schölkopf, Z. Luo and V. Vovk (eds), *Empirical Inference: Festschrift in Honor of Vladimir N. Vapnik*, Springer, pp. 217–230.

Koltchinskii, V. I. (1997). M-estimation, convexity and quantiles, *The Annals of Statistics* **25**(2): 435–477.

Koltchinskii, V., Lounici, K. and Tsybakov, A. B. (2011). Nuclear-norm penalization and optimal rates for noisy low-rank matrix completion, *The Annals of Statistics* **39**(5): 2243–2794.

Kong, E., Linton, O. and Xia, Y. (2010). Uniform Bahadur representation for local polynomial estimates of M-regression and its application to the additive model, *Econometric Theory* **26**(5): 1529–1564.

Kong, L. and Mizera, I. (2012). Quantile tomography: using quantiles with multivariate data, *Statistica Sinica* **22**: 1589–1610.

Kuan, C.-M., Yeh, J.-H. and Hsu, Y.-C. (2009). Assessing value at risk with CARE, the Conditional Autoregressive Expectile models, *Journal of Econometrics* **150**: 261–270.

LaLonde, R. J. (1986). Evaluating the econometric evaluations of training programs with experimental data, *The American Economic Review* **76**(4): 604–620.

Lehmann, E. L. (1975). *Nonparametrics: Statistical Models Based on Ranks*, Springer, San Francisco, CA.

Li, Q., Lin, J. and Racine, J. S. (2013). Optimal bandwidth selection for nonparametric conditional distribution and quantile functions, *Journal of Business & Economic Statistics* **31**(1): 57–65.

Li, Q. and Racine, J. S. (2007). *Nonparametric Econometrics: Theory and Practice*, Princeton university press, New Jersey.

Liu, W. and Wu, W. B. (2010). Simultaneous nonparametric inference of time series, *The Annals of Statistics* **38**(4): 2388–2421.

Lobato, I., Nankervis, J. C. and Savin, N. (2001). Testing for Autocorrelation Using a Modified Box-Pierce Q Test, *International Economic Review* **42**(1): 187–205.

Lounici, K. and Nickl, R. (2011). Global uniform risk bounds for wavelet deconvolution estimators, *The Annals of Statistics* **39**(1): 201–231.

Mammen, E., Van Keilegom, I. and Yu, K. (2013). Expansion for moments of regression quantiles with applications to nonparametric testing, *ArXiv e-prints* .

Meerschaert, M. M., Wang, W. and Xiao, Y. (2013). Fernique-type inequalities and moduli of continuity for anisotropic Gaussian random fields, *Transations of the American Mathematical Society* **365**(2): 1081–1107.

Mojirsheibani, M. (2012). A weighted bootstrap approximation of the maximal deviation of kernel density estimates over general compact sets, *Journal of Multivariate Analysis* **112**: 230–241.

Muhsal, B. and Neumeyer, N. (2010). A note on residual-based empirical likelihood kernel density estimator, *Electronic Journal of Statistics* **4**: 1386–1401.

Negahban, S. N., Ravikumar, P., Wainwright, M. J. and Yu, B. (2012). A unified framework for high-dimensional analysis of M-estimators with decomposable regularizers, *Statistical Science* **27**(4): 538–557.

Negahban, S. N. and Wainwright, M. J. (2011). Estimation of (near) low-rank matrices with nose and high-dimensional scaling, *The Annals of Statistics* **39**(2): 1069–1097.

Nesterov, Y. (2005). Smooth minimization of non-smooth functions, *Mathematical Programming* **103**(1): 127–152.

Newey, W. K. (1997). Convergence rates and asymptotic normality for series estimators, *Journal of Econometrics* **79**: 147–168.

Newey, W. K. and Powell, J. L. (1987). Asymmetric least squares estimation and testing, *Econometrica* **55**(4): 819–847.

Owen, A. B. (2005). *Multidimensional variation for quasi-Monte Carlo*, Vol. 2 of *Ser. Biostat.*, World Sci. Publi., Hackensack, NJ., pp. 49–74.

Proksch, K., Bissantz, N. and Dette, H. (2015). Confidence bands for multivariate and time dependent inverse regression models, *Bernoulli* **21**(1): 144–175.

Reinsel, G. C. and Velu, R. P. (1998). *Multivariate Reduced-Rank Regression*, Springer, New York.

Rosenblatt, M. (1976). On the maximal deviation of k-dimensional density estimates, *The Annals of Probability* **4**(6): 1009–1015.

Ruppert, D., Sheather, S. J. and Wand, M. P. (1995). An effective bandwidth selector for local least squares regression, *Journal of the American Statistical Association* **90**: 1257–1270.

Ruppert, D. and Wand, M. P. (1995). Multivariate locally weighted least squares regression, *The Annal of Statistics* **23**: 1346–1370.

Schaumburg, J. (2011). Predicting extreme VaR: Nonparametric quantile regression with refinements from extreme value theory, *Discussion Paper 2010-009*, CRC 649, Humboldt-Universität zu Berlin.

Schumaker, L. (1981). *Spline Functions: Basic Theory*, Wiley, New York.

Serfling, R. (2002). Quantile functions for multivariate analysis: approaches and applications, *Statistica Neerlandica* **56**(2): 214–232.

Smirnov, N. V. (1950). On the construction of confidence regions for the density of distribution of random variables, *Doklady Akad. Nauk SSSR* **74**: 189–191.

Song, S., Ritov, Y. and Härdle, W. (2012). Partial linear quantile regression and bootstrap confidence bands, *Journal of Multivariate Analysis* **107**: 244–262.

Stone, C. J. (1985). Additive regression and other nonparametric models, *Annals of Statistics* **13**(2): 689–705.

Taylor, J. W. (2008). Using exponentially weighted quantile regression to estimate value at risk and expected shortfall, *Journal of Financial Econometrics* **6**: 382–406.

Toh, K.-C. and Yun, S. (2010). An accelerated proximal gradient algorithm for nuclear norm regularized least squares problems, *Pacific Journal of Optimization* **6**: 615–640.

Tukey, J. W. (1975). Mathematics and picturing data, *in* R. D. James (ed.), *Proceedings of the International Congress on Mathematics*.

Vershynin, R. (2012). *Compressed Sensing, Theory and Applications*, Cambridge University Press, chapter 5, pp. 210–268.

Wainwright, M. J. (2009). Sharp thresholds for high-dimensional and noisy sparsity recovery using ℓ_1-constrained quadratic programming (Lasso), *IEEE Transactions on Information Theory* **55**: 2183–2202.

White, H., Kim, T.-H. and Manganelli, S. (2008). Modeling autoregressive conditional skewness and kurtosis with multi-quantile CAViaR, *in* J. Russell and M. Watson (eds), *Volatility and Time Series Econometrics: A Festschrift in Honor of Robert F. Engle*.

White, H., Kim, T.-H. and Manganelli, S. (2010). VAR for VaR: measuring systemic risk using multivariate regression quantiles, *MPRA Paper No. 35372* .

Yu, K. and Jones, M. C. (1998). Local linear quantile regression, *Journal of the American Statistical Association* **93**(441): 228–237.

Yuan, M., Ekici, A., Lu, Z. and Monteiro, R. (2007). Dimension reduction and coefficient estimation in multivariate linear regression, *Journal of the Royal Statistical Society: Series B* **69**(3): 329–346.

Appendix A

Supplementary materials for Chapter 2

A.1 Locally Linear Quantile Regression (LLQR)

Let $\{(X_i, Y_i)\}_{i=1}^n \subset \mathbb{R}^2$ be i.i.d. bivariate random variables. Denote by $F_{Y|x}(u)$ the conditional cdf and $l(x) = F_{Y|x}^{-1}(\tau)$ the conditional quantile curve to level τ, given observations $\{(x_i, y_i)\}_{i=1}^n$, one may write this as

$$y_i = l(x_i) + \varepsilon_i,$$

with $F_{\varepsilon|x}^{-1}(\tau) = 0$. A locally linear kernel quantile estimator (LLQR) is estimated as $\hat{l}(x_0) = \hat{a}_0$ from:

$$(\hat{a}_0, \hat{b}_0) = \underset{\{a_0, b_0\}}{\operatorname{argmin}} \sum_{i=1}^{n} K\left(\frac{x_i - x_0}{h}\right) \rho_\tau \left\{y_i - a_0 - b_0(x_i - x_0)\right\}, \tag{A.1.1}$$

where h is the bandwidth, $K(\cdot)$ is a kernel and $\rho_\tau(\cdot)$ is the check function given by

$$\rho_\tau(u) = (\tau - \mathbf{1}_{\{u<0\}})u \tag{A.1.2}$$

Figure A.1.1 illustrates the check functions. Different loss functions give different estimates. u^2 corresponds to the conditional mean. $\rho_\tau(u)$ corresponds to the conditional τth quantile.

It is shown by Fan et al. (1994) that the locally linear kernel estimator is asymptotically efficient in a minimax sense. It also possesses good finite sampling property which is adaptive to a variety of empirical density $g(x)$ and has good boundary property.

Next, we describe the method to compute the bandwidths. The approach used here follows Yu and Jones (1998). The bandwidth is chosen by

$$h_\tau = h_{mean} \left[\tau(1-\tau)\varphi\{\Phi^{-1}(\tau)\}^{-2}\right]^{1/5}, \tag{A.1.3}$$

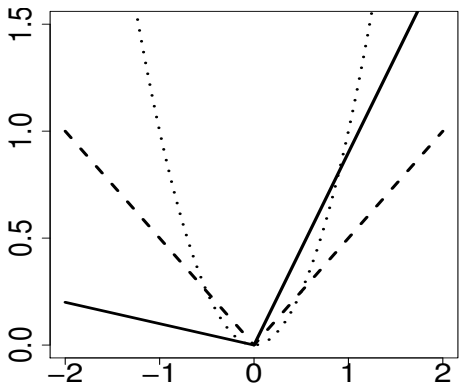


Figure A.1.1: This figure presents the check function. The dotted line is u^2. The dashed and solid lines are check functions $\rho_\tau(u)$ with $\tau = 0.5$ and 0.9 respectively.

where h_{mean} is the locally linear mean regression bandwidth, which can be computed by the algorithm described in Ruppert and Wand (1995) or Ruppert et al. (1995). $\varphi(\cdot)$ and $\Phi(\cdot)$ are the pdf and cdf of the standard normal distribution. Since we discuss the case for VaR, τ is usually small. h_τ needs to be enlarged to allow for more smoothing (usually taking $1.5h_\tau$ or $2h_\tau$).

The approach is acceptable but not so flexible because it is based on assuming the quantile functions are parallel. A more flexible approach was developed by Härdle et al. (2013). In order to stabilize the bandwidth choice, we first regress y_i on the rank of the corresponding x_i and then rescale the resulted estimated values to the original x space. Carroll and Härdle (1989) show that this local bandwidth estimator and the global bandwidth estimator are asymptotically equivalent.

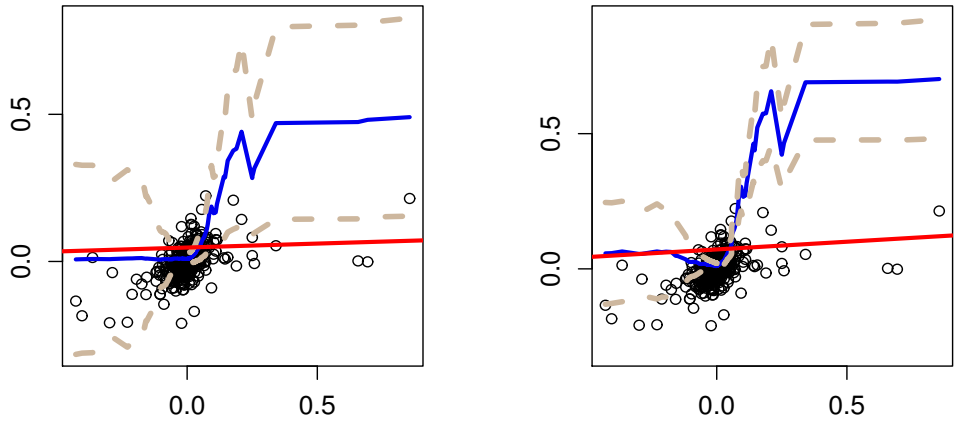


Figure A.1.2: GS and C weekly returns 0.90(left) and 0.95(right) quantile functions. The y-axis is GS daily returns and the x-axis is the C daily returns. The blue curves are the LLQR curves (see Appendix A.1). The LLQR bandwidths are 0.0942 and 0.1026. The red lines are the linear parametric quantile regression line. The antique white curves are the asymptotic confidence band (see Appendix A.2) with significance level 0.05. $n = 546$.

A.2 Confidence band for nonparametric quantile estimator

The uniform confidence band of the quantile estimator is based on the Theorem 2.2 and Corollary 2.1 presented in Härdle and Song (2010). The details are as follows.

Let $\{(X_i, Y_i)\}_{i=1}^n$ be as in Appendix A.1. Define $K_h(u) = h^{-1}K(u/h)$ and similar to (A.1.1) let $l_n(x)$ and $l(x)$ are zeros (w.r.t. θ) of the functions:

$$\tilde{H}_n(\theta, x) \stackrel{\text{def}}{=} n^{-1} \sum_{i=1}^{n} K_h(x - X_i)\rho_\tau(Y_i - \theta);$$
$$\tilde{H}(\theta, x) \stackrel{\text{def}}{=} \int_{\mathbb{R}} f(x, y)\rho_\tau(y - \theta)dy,$$

where $\rho_\tau(\cdot)$ is the check function defined as (A.1.2).

THEOREM A.2.1. Let $h = n^{-\delta}$, $\frac{1}{5} < \delta < \frac{1}{3}$, $\lambda(K) = \int_{-A}^{A} K^2(u)du$, where $K(\cdot)$ is supported on $[-A, A]$. $J = [0, 1]$. Define $c_1(K) = \{K^2(A) + K^2(-A)\}/2\lambda(K)$, $c_2(K) = \int_{-A}^{A}\{K'(u)\}^2 du/2\lambda(K)$ and

$$d_n = \begin{cases} (2\delta \log n)^{1/2} + (2\delta \log n)^{-1/2}\left[\log\{c_1(K)/\pi^{1/2}\} + \frac{1}{2}\{\log\delta + \log\log n\}\right], \\ \qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad \text{if } c_1(K) > 0; \\ (2\delta \log n)^{1/2} + (2\delta \log n)^{-1/2}\log\{c_2(K)/2\pi\}, \text{ otherwise.} \end{cases}$$

Then

$$\mathrm{P}\left[(2\delta\log n)^{1/2}\left\{\sup_{x\in J} r(x)|l_n(x)-l(x)|/\lambda(K)^{1/2}-d_n\right\}<z\right]\to\exp\{-2\exp(-z)\},$$

as $n\to\infty$, with

$$r(x)=(nh)^{1/2}f\{l(x)|x\}\{f_X(x)/\tau(1-\tau)\}^{1/2},$$

where $f_X(\cdot)$ is the marginal pdf for X and $f(\cdot|x)$ is the conditional pdf of Y on $X=x$.

The corollary followed by the theorem explicitly indicates how a uniform confidence interval can be constructed.

COROLLARY A.2.2. An approximate $(1-\alpha)\times 100\%$ confidence band is

$$l_n\pm(nh)^{-1/2}\{\tau(1-\tau)\lambda(K)/\hat{f}_X(t)\}^{1/2}\hat{f}^{-1}\{l(t)|t\}\{d_n+c(\alpha)(2\delta\log n)^{-1/2}\},$$

where $c(\alpha)=\log 2-\log|\log(1-\alpha)|$ and $\hat{f}_X(t)$, $\hat{f}\{l(t)|t\}$ are consistent estimates for $f_X(t)$, $f\{l(t)|t\}$.

Figure 2.1.1 is done by the techniques introduced in Appendices A.1 and A.2. Another illustration with right tail quantiles is in Figure A.1.2. We plot the LLQR curve for 0.9 and 0.95 quantile. Both the two linear quantile regression lines lie outside the LLQR confidence band as the Citigroup returns are positive.

A.3 PLM model estimation

For the PLM estimation, we adopt the algorithm described in Song et al. (2012). Given data $\{(X_t,Y_t)\}_{t=1}^T$ bivariate and $\{M_t\}_{t=1}^T$ multivariate random variables. The PLM is:

$$y_t=\alpha+\beta^\top M_{t-1}+l(x_t)+\varepsilon_t.$$

Let a_n denote an increasing sequence of positive integers and set $b_n=a_n^{-1}$. For each $n=1,2,...$, dividing the interval $[0,1]$ in a_n subintervals I_{nt}, $t=1,...,a_n$ with equal length b_n. On each I_{nt}, $l(\cdot)$ can approximately be taken as a constant.

The PLM estimation procedure is:

1. Inside each partition I_{nt}, a linear quantile regression is performed to get $\hat{\beta}_i$, then their weighted mean gives $\hat{\beta}$. Formally, let $\rho_\tau(\cdot)$ be the check function defined as (A.1.2), $l_1,...,l_{a_n}$ are constants,

$$\hat{\beta}=\operatorname*{argmin}_{\beta}\min_{l_1,...,l_{a_n}}\sum_{t=1}^{n}\rho_\tau\left\{X_{j,t}-\alpha-\beta^\top M_{t-1}-\sum_{m=1}^{a_n}l_m\mathbf{1}(X_{i,t}\in I_{nt})\right\}$$

2. Computing the LLQR nonparametric quantile estimates of $l(\cdot)$ as outlined in Appendix A.1 from $\{(X_{i,t},X_{j,t}-\hat{\alpha}-\hat{\beta}^\top M_{t-1})\}_{t=1}^N$.

Appendix B

Supplementary materials for Chapter 3

B.1 Proof of Theorems

We list the assumptions here for the easy of reference.

(A1) K is of order $s-1$ (see (A3)) has bounded support $[-A, A]^d$, is continuously differentiable up to order d with bounded derivatives, i.e. $\partial^{\boldsymbol{\alpha}} K \in L^1(\mathbb{R}^d)$ exists and is continuous for all multi-indices $\boldsymbol{\alpha} \in \{0,1\}^d$

(A2) Let a_n be an increasing sequence, $a_n \to \infty$ as $n \to \infty$, and the marginal density f_Y be such that

$$(\log n) h^{-3d} \int_{|y|>a_n} f_Y(y) dy = \mathcal{O}(1) \tag{B.1.1}$$

and

$$(\log n) h^{-d} \int_{|y|>a_n} f_{Y|\boldsymbol{X}}(y|\boldsymbol{x}) dy = \mathcal{O}(1), \text{ for all } \boldsymbol{x} \in \mathcal{D}$$

as $n \to \infty$ hold.

(A3) The function $\theta_0(\boldsymbol{x})$ is continuously differentiable and is in Hölder class with order $s > d$.

(A4) $f_{\boldsymbol{X}}(\boldsymbol{x})$ is bounded, continuously differentiable and its gradient is uniformly bounded. Moreover, $\inf_{\boldsymbol{x}\in\mathcal{D}} f_{\boldsymbol{X}}(\boldsymbol{x}) > 0$.

(A5) The joint probability density function $f(y, \boldsymbol{u})$ is bounded, positive and continuously differentiable up to sth order (needed for Rosenblatt transform). The conditional density $f_{Y|\boldsymbol{X}}(y|\boldsymbol{x})$ exists and is boudned and continuouly differentiable with respect to $\boldsymbol{x}$. Moreover, $\inf_{\boldsymbol{x}\in\mathcal{D}} f_{Y|\boldsymbol{X}}\big(\theta_0(\boldsymbol{x})|\boldsymbol{x}\big) > 0$.

(A6) h satisfies $\sqrt{nh^d} h^s \sqrt{\log n} \to 0$ (undersmoothing), and $nh^{3d}(\log n)^{-2} \to \infty$.

(EA2) $\sup_{\boldsymbol{x}\in\mathcal{D}}\left|\int v^{b_1} f_{\varepsilon|\boldsymbol{X}}(v|\boldsymbol{x})dv\right| < \infty$, for some $b_1 > 0$.

(B1) $L \in L^1(\mathbb{R}^d)$ is a Lipschitz, bounded, symmetric kernel. G is Lipschitz continuous cdf with $G(x), 1-G(x) \leq Ce^{-x}$ for $C > 0$, and $g \in L^1(\mathbb{R})$ is the derivative of G and is also a density, which is Lipschitz continuous, bounded, symmetric and five times continuously differentiable kernel.

(B2) $F_{\varepsilon|\boldsymbol{X}}(v|\boldsymbol{x})$ is in $s'+1$ order Hölder class with respect to v and continuous in $\boldsymbol{x}$, $s' > \max\{2, d\}$. $f_{\boldsymbol{X}}(\boldsymbol{x})$ is in second order Hölder class with respect to $\boldsymbol{x}$ and v. $\mathsf{E}[\psi^2(\varepsilon_i)|\boldsymbol{x}]$ is second order continuously differentiable with respect to $\boldsymbol{x} \in \mathcal{D}$.

(B3) $nh_0\bar{h}^d \to \infty$, $h_0, \bar{h} = \mathcal{O}(n^{-\nu})$, where $\nu > 0$.

(C1) There exist an increasing sequence c_n, $c_n \to \infty$ as $n \to \infty$ such that

$$(\log n)^3 (nh^{6d})^{-1} \int_{|v|>c_n/2} f_\varepsilon(v)dv = \mathcal{O}(1), \tag{B.1.2}$$

as $n \to \infty$.

(EC1) $\sup_{\boldsymbol{x}\in\mathcal{D}}\left|\int |v|^b f_{\varepsilon|\boldsymbol{X}}(v|\boldsymbol{x})dv\right| < \infty$, for some $b > 0$.

Define the approximating processes

$$Y_n(\boldsymbol{x}) \stackrel{\text{def}}{=} \frac{1}{\sqrt{h^d f_{\boldsymbol{X}}(\boldsymbol{x})}\sigma(\boldsymbol{x})} \int\int K\left(\frac{\boldsymbol{x}-\boldsymbol{u}}{h}\right)\psi_\tau(y-\theta_0(\boldsymbol{x}))dZ_n(y,\boldsymbol{u}). \tag{B.1.3}$$

$$Y_{0,n}(\boldsymbol{x}) = \frac{1}{\sqrt{h^d f_{\boldsymbol{X}}(\boldsymbol{x})\sigma_n^2(\boldsymbol{x})}} \int\int_{\Gamma_n} K\left(\frac{\boldsymbol{x}-\boldsymbol{u}}{h}\right)\psi_\tau(y-\theta_0(\boldsymbol{x}))dZ_n(y,\boldsymbol{u}), \tag{B.1.4}$$

where $\Gamma_n = \{y : |y| \leq a_n\}$ and $\sigma_n^2(\boldsymbol{x}) = \mathsf{E}[\psi^2(Y-\theta_0(\boldsymbol{x}))\mathbf{1}(Y_i \leq a_n)|\boldsymbol{X} = \boldsymbol{x}]$.

$$Y_{1,n}(\boldsymbol{x}) = \frac{1}{\sqrt{h^d f_{\boldsymbol{X}}(\boldsymbol{x})\sigma_n^2(\boldsymbol{x})}} \int\int_{\Gamma_n} K\left(\frac{\boldsymbol{x}-\boldsymbol{u}}{h}\right)\psi_\tau(y-\theta_0(\boldsymbol{x}))dB_n\big(T(y,\boldsymbol{u})\big) \tag{B.1.5}$$

where $B_n\big\{T(y,\boldsymbol{u})\big\} = W_n\big\{T(y,\boldsymbol{u})\big\} - F(y,\boldsymbol{u})W_n(1,...,1)$ and $T(y,\boldsymbol{u})$ is the Rosenblatt transformation

$$T(y,\boldsymbol{u}) = \big\{F_{X_1|Y}(u_1|y), F_{X_2|Y}(u_2|u_1,y), ..., F_{X_d|X_{d-1},...,X_1,Y}(u_d|u_{d-1},...,u_1,y), F_Y(y)\big\}.$$

$$Y_{2,n}(\boldsymbol{x}) = \frac{1}{\sqrt{h^d f_{\boldsymbol{X}}(\boldsymbol{x})\sigma_n^2(\boldsymbol{x})}} \int\int_{\Gamma_n} K\left(\frac{\boldsymbol{x}-\boldsymbol{u}}{h}\right)\psi_\tau(y-\theta_0(\boldsymbol{x}))dW_n\big(T(y,\boldsymbol{u})\big) \tag{B.1.6}$$

$$Y_{3,n}(\boldsymbol{x}) = \frac{1}{\sqrt{h^d f_{\boldsymbol{X}}(\boldsymbol{x})\sigma_n^2(\boldsymbol{x})}} \int\int_{\Gamma_n} K\left(\frac{\boldsymbol{x}-\boldsymbol{u}}{h}\right) \psi_\tau(y-\theta_0(\boldsymbol{u}))dW_n\big(T(y,\boldsymbol{u})\big) \tag{B.1.7}$$

$$Y_{4,n}(\boldsymbol{x}) = \frac{1}{\sqrt{h^d f_{\boldsymbol{X}}(\boldsymbol{x})\sigma_n^2(\boldsymbol{x})}} \int \sqrt{\sigma_n^2(\boldsymbol{u}) f_{\boldsymbol{X}}(\boldsymbol{u})} K\left(\frac{\boldsymbol{x}-\boldsymbol{u}}{h}\right) dW\big(\boldsymbol{u}\big). \tag{B.1.8}$$

$$Y_{5,n}(\boldsymbol{x}) = \frac{1}{\sqrt{h^d}} \int K\left(\frac{\boldsymbol{x}-\boldsymbol{u}}{h}\right) dW\big(\boldsymbol{u}\big). \tag{B.1.9}$$

In these approximating processes, the function

$$\psi_\tau(u) = \begin{cases} \mathbf{1}(u \le 0) - \tau, & \text{Quantile;} \\ 2(\mathbf{1}(u \le 0) - \tau)|u|, & \text{Expectile.} \end{cases}$$

In the proofs, we suppress the subscript "τ".

Next we introduce some notations which are used repeatedly in the following proofs.

Definition B.1.1 (Neighboring Block in $\mathcal{D} \subset \mathbb{R}^d$, Bickel and Wichura (1971) p.1658)**.** A *block* $B \subset \mathcal{D}$ is a subset of $\mathcal{D}$ of the form $B = \Pi_i(s_i, t_i]$ with s and t in $\mathcal{D}$; the *pth-face* of B is $\Pi_{i\neq p}(s_i, t_i]$. Disjoint blocks B and C are *p-neigbbors* if they abut and have the same pth face; they are *neighbors* if they are p-neighbors for some $p \ge 1$.

To illustrate the idea of neighboring block, take $d = 3$ for example, the blocks $(s,t] \times (a,b] \times (c,d]$ and $(t,u] \times (a,b] \times (c,d]$ are 1-neighbors for $s \le t \le u$.

Definition B.1.2 (Bickel and Wichura (1971) p.1658)**.** Let $X : \mathbb{R}^d \to \mathbb{R}$. The *increment of* X *on the block* B, denoted $X(B)$, is defined by

$$X(B) = \sum_{\boldsymbol{\alpha}\in\{0,1\}^d} (-1)^{d-|\boldsymbol{\alpha}|} X\big\{\boldsymbol{s} + \boldsymbol{\alpha} \odot (\boldsymbol{t} - \boldsymbol{s})\big\}, \tag{B.1.10}$$

where "$\odot$" denotes the *componentwise product*; that is, for any vectors $\boldsymbol{u}, \boldsymbol{v} \in \mathbb{R}^d$, $\boldsymbol{u} \odot \boldsymbol{v} = (u_1 v_1, u_2 v_2, ..., u_d v_d)$.

Below we give some examples of the increment of a multivariate function X on a block:

- $d = 1$: $B = (s,t]$, $X(B) = X(t) - X(s)$;
- $d = 2$: $B = (s_1, t_1] \times (s_2, t_2]$. $X(B) = X(t_1,t_2) - X(t_1,s_2) + X(s_1,s_2) - X(s_1,t_2)$.

B.1.1 Proof of Theorem 3.2.1

LEMMA B.1.3.

$$\|Y_n(\boldsymbol{x}) - Y_{0,n}(\boldsymbol{x})\| = \mathcal{o}_p\big\{(\log n)^{-1/2}\big\},$$

where $\|\cdot\|$ denotes the sup norm with respect to $\boldsymbol{x} \in \mathcal{D}$.

PROOF. By the triangle inequality we have

$$\|Y_n - Y_{n,0}\| \leq \|Y_n - \hat{Y}_{n,0}\| + \|\hat{Y}_{n,0} - Y_{n,0}\| \stackrel{\text{def}}{=} E_1 + E_2,$$

where $\hat{Y}_{n,0} = \sigma^2(\boldsymbol{x})/\sigma_n(\boldsymbol{x})Y_{n,0}(\boldsymbol{x})$ and the terms E_1 and E_2 are defined in an obvious manner. We now show that $E_j = \mathcal{o}_p\big\{(\log n)^{-1/2}\big\}$, $j = 1, 2$. Note that

$$|\hat{Y}_{n,0}(\boldsymbol{x}) - Y_{n,0}(\boldsymbol{x})| = \Big|\big(\sigma(\boldsymbol{x})/\sigma_n(\boldsymbol{x}) - 1\big)Y_{n,0}(\boldsymbol{x})\Big|.$$

It is shown later that $\|Y_{n,0}\| = \mathcal{O}_p\big(\sqrt{\log n}\big)$, hence it remains to prove that

$$\sup_{\boldsymbol{x}\in\mathcal{D}} \big|\sigma(\boldsymbol{x})/\sigma_n(\boldsymbol{x}) - 1\big| = \mathcal{o}\big\{(\log n)^{-1}\big\}. \tag{B.1.11}$$

To this end let $\tilde{\sigma}_n^2 = \mathsf{E}[\psi^2\{Y_i - \theta_0(\boldsymbol{x})\}\mathbf{1}(|Y_i| > a_n)|\boldsymbol{X} = \boldsymbol{x}]$. Since $\sigma_n^2(\boldsymbol{x}) \to \tau(1-\tau) > 0$ for $n \to \infty$, by (B.1.1), and $\psi^2(\cdot) \leq \max\{\tau^2, (1-\tau)^2\}$, $|(\log n)^2\tilde{\sigma}_n^2(\boldsymbol{x})/\sigma_n^2(\boldsymbol{x})| \leq \big|(\log n)h^d\mathcal{O}(1)\big| \to 0$. Therefore,

$$\begin{aligned}(\log n)\sup_{\boldsymbol{x}\in\mathcal{D}}\left|\sqrt{\frac{\sigma^2(\boldsymbol{x})}{\sigma_n^2(\boldsymbol{x})}} - 1\right| &= (\log n)\sup_{\boldsymbol{x}\in\mathcal{D}}\left|\sqrt{\frac{\tilde{\sigma}_n^2(\boldsymbol{x}) + \sigma_n^2(\boldsymbol{x})}{\sigma_n^2(\boldsymbol{x})}} - 1\right| \\ &\leq \sup_{\boldsymbol{x}\in\mathcal{D}}\left|\sqrt{\frac{(\log n)^2\tilde{\sigma}_n^2(\boldsymbol{x})}{\sigma_n^2(\boldsymbol{x})}}\right| \to 0,\end{aligned}$$

as $n \to \infty$, hence $E_2 = \mathcal{o}_p\big((\log n)^{-1/2}\big)$. We now use Lemma B.2.2 in order to show that E_1 too is negligible.

$$\begin{aligned}&(\log n)^{1/2}E_1 \\ &= (\log n)^{1/2}\sup_{\boldsymbol{x}\in\mathcal{D}}|Y_n(\boldsymbol{x}) - \hat{Y}_{n,0}(\boldsymbol{x})| \\ &= (\log n)^{1/2}\sup_{\boldsymbol{x}\in\mathcal{D}}\left|\frac{1}{\sqrt{h^d f_{\boldsymbol{X}}(\boldsymbol{x})\sigma^2(\boldsymbol{x})}}\int\int_{\{|y|>a_n\}} K\left(\frac{\boldsymbol{x}-\boldsymbol{u}}{h}\right)\psi\{y - \theta_0(\boldsymbol{x})\}dZ_n(y,\boldsymbol{u})\right| \\ &= \sup_{\boldsymbol{x}\in\mathcal{D}}\left|\frac{1}{\sqrt{f_{\boldsymbol{X}}(\boldsymbol{x})\sigma^2(\boldsymbol{x})}}V_n(\boldsymbol{x})\right|,\end{aligned}$$

where

$$V_n(\boldsymbol{x}) = \sum_{i=1}^n W_{n,i}(\boldsymbol{x}),$$

and

$$W_{n,i}(\boldsymbol{x}) = (\log n)^{1/2}(nh^d)^{-1/2}\left\{\psi(Y_i - \theta_0(\boldsymbol{x}))\mathbf{1}(|Y_i| > a_n)K\Big(\frac{\boldsymbol{x}-\boldsymbol{X}_i}{h}\Big)\right.$$
$$\left. -\mathsf{E}\Big[\psi(Y_i - \theta_0(\boldsymbol{x}))\mathbf{1}(|Y_i| > a_n)K\Big(\frac{\boldsymbol{x}-\boldsymbol{X}_i}{h}\Big)\Big]\right\}.$$

Note that $f_{\boldsymbol{X}}(\boldsymbol{x})\sigma^2(\boldsymbol{x}) = f_{\boldsymbol{X}}(\boldsymbol{x})\tau(1-\tau) > 0$ for $\boldsymbol{x} \in \mathcal{D}$ by Assumption (A4).

$$\mathsf{E}[W_{n,i}(\boldsymbol{x})^2] \le (\log n)(nh^d)^{-1}\mathsf{E}\left[\psi^2(Y_i - \theta_0(\boldsymbol{x}))\mathbf{1}(|Y_i| > a_n)K^2\Big(\frac{\boldsymbol{x}-\boldsymbol{X}_i}{h}\Big)\right]$$
$$\le (\log n)(nh^d)^{-1}C_{\psi,K}\int_{\{|y|>a_n\}} f_Y(y)dy.$$

Thus, from (B.1.1),

$$\mathsf{E}\Big[\big(\sum_{i=1}^n W_{n,i}(\boldsymbol{x})\big)^2\Big] \le (\log n)h^{-d}C_{\psi,K}\int_{\{|y|>a_n\}} f_Y(y)dy = h^{2d}\mathcal{O}_p(1) \to 0,$$

as $n \to \infty$. From Markov's inequality, $|V_n(\boldsymbol{x})| \xrightarrow{p} 0$ for each fixed $\boldsymbol{x} \in \mathcal{D}$.

We now show the tightness of $V_n(\boldsymbol{x})$ for $\boldsymbol{x} \in \mathcal{D}$ in order to obtain the uniform convergence. To simplify the expression, define

$$g(\boldsymbol{x}) \stackrel{\text{def}}{=} \psi\{y - \theta_0(\boldsymbol{x})\}K\left(\frac{\boldsymbol{x}-\boldsymbol{u}}{h}\right).$$

Take arbitrary neighboring blocks $B, C \subset \mathcal{D}$ (see Definition B.1.1) and suppose $B = \Pi_{i=1}^d(s_i, t_i]$,

$$\mathsf{E}[V_n(B)^2]^{1/2} \le (\log n)^{1/2}h^{-d/2}\Big\{\mathsf{E}\Big[\mathbf{1}(Y_i > a_n)\Big(\sum_{\boldsymbol{\alpha}\in\{0,1\}^d}(-1)^{d-|\boldsymbol{\alpha}|}g\big(\boldsymbol{s}+\boldsymbol{\alpha}\odot(\boldsymbol{t}-\boldsymbol{s})\big)\Big)^2\Big]$$
$$+\mathsf{E}\Big[\mathbf{1}(Y_i < -a_n)\Big(\sum_{\boldsymbol{\alpha}\in\{0,1\}^d}(-1)^{d-|\boldsymbol{\alpha}|}g\big(\boldsymbol{s}+\boldsymbol{\alpha}\odot(\boldsymbol{t}-\boldsymbol{s})\big)\Big)^2\Big]\Big\}^{1/2}$$
$$\stackrel{\text{def}}{=} (\log n)^{1/2}h^{-d/2}(I_1+I_2)^{1/2},$$

where I_1 and I_2 are defined in an obvious manner. When n is large, a_n is large as well and the integral is restricted to the set $\{Y_i > a_n\}$. Taking into account that θ is uniformly bounded on the compact set $\mathcal{D}$ by Assumption (A4) we deduce that $\psi(Y_i - \theta_0(\boldsymbol{x})) = \tau$ for sufficiently large n on the event $\{Y_i > a_n : i = 1, ..., n\}$. Hence, I_1 can be estimated as

$$I_1$$
$$\le \tau^2\int\int \mathbf{1}(y > a_n)\Big(\sum_{\boldsymbol{\alpha}\in\{0,1\}^d}(-1)^{d-|\boldsymbol{\alpha}|}K\Big[\big(\boldsymbol{s}+\boldsymbol{\alpha}\odot(\boldsymbol{t}-\boldsymbol{s})-\boldsymbol{u}\big)/h\Big]\Big)^2 f(y,\boldsymbol{u})dyd\boldsymbol{u}.$$

Note that

$$\sum_{\boldsymbol{\alpha}\in\{0,1\}^d}(-1)^{d-|\boldsymbol{\alpha}|}K\Big[\big(\boldsymbol{s}+\boldsymbol{\alpha}\odot(\boldsymbol{t}-\boldsymbol{s})-\boldsymbol{u}\big)/h\Big]$$
$$=\int_B \partial^{(1,\ldots,1)}K\Big(\frac{\boldsymbol{v}-\boldsymbol{u}}{h}\Big)d\boldsymbol{v}\le h^{-d}C_{K'}\mu(B),$$

where the constant $C_{K'}$ satisfies $\sup_{\boldsymbol{u}\in\mathcal{D}}|\partial^{\boldsymbol{\alpha}}K(\boldsymbol{u})|\le C_{K'}$ and $\mu(\cdot)$ is the Lebesgue measure. As consequence it follows that

$$\begin{aligned}I_1&\le\tau^2\int\int\mathbf{1}(y>a_n)\big(C_{K'}\mu(B)\big)^2f(y,\boldsymbol{u})dyd\boldsymbol{u}\\&=\tau^2\big(h^{-d}C_{K'}\mu(B)\big)^2\int_{\{y>a_n\}}f_Y(y)dy.\end{aligned}$$

Similarly, $I_2\le(1-\tau)^2\big(C_{K'}h^{-d}\mu(B)\big)^2\int_{\{y<-a_n\}}f_Y(y)dy$. Hence,

$$\begin{aligned}&\mathsf{E}[V_n(B)^2]^{1/2}\\&\le(\log n)^{1/2}h^{-3d/2}C_{K'}\mu(B)\left(\tau^2\int_{\{y>a_n\}}f_Y(y)dy+(1-\tau)^2\int_{\{y<-a_n\}}f_Y(y)dy\right)^{1/2}\\&\le(\log n)^{1/2}h^{-3d/2}C_{K'}\max(\tau,1-\tau)\left(\int_{\{|y|>a_n\}}f_Y(y)dy\right)^{1/2}\mu(B).\end{aligned}$$

Analogously we obtain the estimate

$$\mathsf{E}[V_n(C)^2]^{1/2}\le(\log n)^{1/2}h^{-3d/2}C_{K'}\max(\tau,1-\tau)\left(\int_{\{|y|>a_n\}}f_Y(y)dy\right)^{1/2}\mu(C),$$

which finally yields

$$\begin{aligned}\mathsf{E}[|V_n(B)||V_n(C)|]&\le\mathsf{E}[|V_n(B)|^2]^{1/2}\mathsf{E}[|V_n(C)|^2]^{1/2}\\&\le(\log n)h^{-3d}C_{K'}^2\max(\tau,1-\tau)^2\left(\int_{\{|y|>a_n\}}f_Y(y)dy\right)\mu(C)\mu(B).\end{aligned}$$

By Assumption (A2) it follows $(\log n)h^{-3d}\int_{\{|y|>a_n\}}f_Y(y)dy$ is bounded. Thus, applying Lemma B.2.2 with $\gamma_1=\gamma_2=\lambda_1=\lambda_2=1$ yields the desired result. □

LEMMA B.1.4. $\|Y_{0,n}-Y_{1,n}\|=\mathcal{O}_p\big(n^{-1/6}h^{-d/2}(\log n)^{\epsilon+(2d+4)/3}\big)$, a.s. for any $\epsilon>0$.

PROOF. We adopt the notation that if $\boldsymbol{\alpha}\in\{0,1\}^{d+1}$, then we write $\boldsymbol{\alpha}=(\alpha_1,\boldsymbol{\alpha}_2)$ where $\alpha_1\in\{0,1\}$ and $\boldsymbol{\alpha}_2\in\{0,1\}^d$. In the computation below, we focus on $B_{\boldsymbol{x}}=\Pi_{j=1}^d\big[x_j-Ah,x_j+Ah\big]$ instead of $\mathbb{R}^d$ since K has compact support. Recall definition

B.1.1 of an increment of a function X over a block B. Integration by parts yields

$$
\begin{aligned}
&Y_{0,n}(\boldsymbol{x}) && \text{(B.1.12)}\\
&= \frac{1}{\sqrt{h^d f_{\boldsymbol{X}}(\boldsymbol{x})\sigma_n^2(\boldsymbol{x})}}\Big[\int_{B_{\boldsymbol{x}}}\int_{\Gamma_n} Z_n(y,\boldsymbol{u})\; d\big(\psi(y-\theta_0(\boldsymbol{x}))K((\boldsymbol{x}-\boldsymbol{u})/h)\big)\\
&+\Big\{Z_n\big(\cdot_1,\cdot_2\big)\psi\big(\,\cdot_1-\theta_0(\boldsymbol{x})\big)K\Big(\frac{\boldsymbol{x}-\cdot_2}{h}\Big)\Big\}\big(\Gamma_n\times B_{\boldsymbol{x}}\big) && \text{(B.1.13)}\\
&+\Big\{\sum_{\substack{\boldsymbol{\alpha}\in\{0,1\}^{d+1}\\ \boldsymbol{\alpha}\neq(\boldsymbol{0},\boldsymbol{1})}}\int\int_{(\Gamma_n\times B_{\boldsymbol{x}})_{\boldsymbol{\alpha}}} Z_n(\cdot_1,\cdot_2)\; d^{\alpha_1}\psi\big(\,\cdot_1-\theta_0(\boldsymbol{x})\big)\partial^{\boldsymbol{\alpha}_2}K\big((\boldsymbol{x}-\cdot_2)/h\big)\Big\}\\
&\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\big(\Gamma_n\times B_{\boldsymbol{x}}\big)_{\boldsymbol{1}-\boldsymbol{\alpha}}\Big]
\end{aligned}
$$

where $\boldsymbol{1}=(1,...,1)\in\{0,1\}^{d+1}$ and $\boldsymbol{0}=(0,...,0)\in\{0,1\}^{d+1}$. $\big(\Gamma_n\times B_{\boldsymbol{x}}\big)$ is a $d+1$ dimensional cube. $\cdot_1$ corresponds to the one-dimensional variable y and $\cdot_2$ corresponds to the two-dimensional variable u. The second term in (B.1.13) can be evaluated with the formula (B.1.10). $\big(\Gamma_n\times B_{\boldsymbol{x}}\big)_{\boldsymbol{1}-\boldsymbol{\alpha}}$ can be viewed as the projection of $\Gamma_n\times B_{\boldsymbol{x}}$ on to the space spanned by those axes whose numbers correspond to positions of ones of the multi-index $\boldsymbol{1}-\boldsymbol{\alpha}$. This leaves us with an $|\boldsymbol{\alpha}|$-fold integral.

Moreover, $d\big\{\psi(y-\theta_0(\boldsymbol{x}))K((\boldsymbol{x}-\boldsymbol{u})/h)\big\}=d\psi(y-\theta_0(\boldsymbol{x}))\partial^{\boldsymbol{1}_2}K\big((\boldsymbol{x}-\boldsymbol{u})/h\big)$, where $\boldsymbol{1}_2=(1,...,1)\in\{0,1\}^d$ and $d\psi(y-\theta_0(\boldsymbol{x}))=\delta_{\theta_0(\boldsymbol{x})}(y)$ denotes the Dirac measure at $\theta_0(\boldsymbol{x})$.

By integration by parts applied to $Y_{1,n}$ and an application of Theorem 3.2 in Dedecker et al. (2014) we obtain for every $\epsilon>0$, it holds almost surely that

$$
\begin{aligned}
&h^{d/2}n^{1/6}(\log n)^{-\epsilon-(2d+4)/3}|Y_{0,n}-Y_{1,n}|\\
&\le \mathcal{O}(1)\Big|\frac{1}{\sqrt{f_{\boldsymbol{X}}(\boldsymbol{x})\sigma_n^2(\boldsymbol{x})}}\Big|\Big\{\Big|\int_{B_{\boldsymbol{x}}} dK((\boldsymbol{x}-\boldsymbol{u})/h)\Big|\\
&+\Big|\Big\{\psi\big(\,\cdot_1-\theta_0(\boldsymbol{x})\big)K\Big(\frac{\boldsymbol{x}-\cdot_2}{h}\Big)\Big\}\big(\Gamma_n\times B_{\boldsymbol{x}}\big)\Big|\\
&+\Big|\sum_{\alpha_1=1,\boldsymbol{\alpha}_2\in\{0,1\}^d-\{\boldsymbol{1}\}}\int_{(B_{\boldsymbol{x}})_{\boldsymbol{\alpha}_2}}\partial^{\boldsymbol{\alpha}_2}K\big((\boldsymbol{x}-\cdot_2)/h\big)\Big|(B_{\boldsymbol{x}})_{\boldsymbol{1}_2-\boldsymbol{\alpha}_2}\\
&+\Big|\sum_{\alpha_1=0,\boldsymbol{\alpha}_2\in\{0,1\}^d-\{\boldsymbol{0}\}}\int_{(B_{\boldsymbol{x}})_{\boldsymbol{\alpha}_2}}\partial^{\boldsymbol{\alpha}_2}K\big((\boldsymbol{x}-\cdot_2)/h\big)\Big||\psi\big(\,\cdot_1-\theta_0(\boldsymbol{x})\big)|\big(\Gamma_n\times(B_{\boldsymbol{x}})_{\boldsymbol{1}_2-\boldsymbol{\alpha}_2}\big)\Big\}
\end{aligned}
\qquad \text{(B.1.14)}
$$

By (A1), K is of bounded variation in the sense of Hardy and Krause (Owen (2005) definition 2), and this leads to the desired result that (B.1.14) is bounded. □

LEMMA B.1.5. $\|Y_{1,n}-Y_{2,n}\|=\mathcal{O}_p\big(h^{d/2}\big)$.

PROOF. Since $B_n\big(T(y,\boldsymbol{u})\big)=W_n\big(T(y,\boldsymbol{u})\big)-F(y,\boldsymbol{u})W(1,...,1)$, where $T(y,\boldsymbol{u})$ is the Rosenblatt transformation and the Jacobian of $T(y,\boldsymbol{u})$ is $f(y,\boldsymbol{u})$,by a change of

variables and the first order approximation to $f(y, \boldsymbol{x} - h\boldsymbol{v})$:

$$
\begin{aligned}
&|Y_{1,n}(\boldsymbol{x}) - Y_{2,n}(\boldsymbol{x})| \\
&\leq \left| \frac{1}{\sqrt{h^d f_{\boldsymbol{X}}(\boldsymbol{x}) \sigma_n^2(\boldsymbol{x})}} \int \int_{\Gamma_n} K\left(\frac{\boldsymbol{x} - \boldsymbol{u}}{h}\right) \psi(y - \theta_0(\boldsymbol{x})) f(y, \boldsymbol{u}) dy d\boldsymbol{u} \right| |W(1, ..., 1)| \\
&\leq \left| \frac{1}{\sqrt{h^d f_{\boldsymbol{X}}(\boldsymbol{x}) \sigma_n^2(\boldsymbol{x})}} \int \int_{\Gamma_n} K(\boldsymbol{v}) \, \psi(y - \theta_0(\boldsymbol{x})) f(y, \boldsymbol{x} - h\boldsymbol{v}) h^d dy d\boldsymbol{v} \right| |W(1, ..., 1)| \\
&\leq h^{d/2} \left| \int K(\boldsymbol{v}) \, d\boldsymbol{v} \right| \left| \frac{1}{\sqrt{f_{\boldsymbol{X}}(\boldsymbol{x}) \sigma_n^2(\boldsymbol{x})}} \int_{\Gamma_n} |\psi(y - \theta_0(\boldsymbol{x}))| f(y, \boldsymbol{x}) dy + \mathcal{O}(h) \right| \\
&\qquad\qquad |W(1, ..., 1)| \\
&\leq h^{d/2} \left| \int K(\boldsymbol{v}) \, d\boldsymbol{v} \right| \left| \frac{1}{\sqrt{f_{\boldsymbol{X}}(\boldsymbol{x}) \sigma_n^2(\boldsymbol{x})}} \max\{\tau, 1 - \tau\} + \mathcal{O}(h) \right| |W(1, ..., 1)|,
\end{aligned}
$$

note that $|W(1, ..., 1)| = \mathcal{O}_p(1)$. □

LEMMA B.1.6. $\|Y_{2,n} - Y_{3,n}\| = \mathcal{O}_p\big(h^{1/2-\delta}\big)$ for an arbitrarily small $0 < \delta < 1/2$.

REMARK B.1.7. We note that the rate of $h^{1/2-\delta}$ is not sharp rate but sufficiently fast for our purpose.

PROOF. Define

$$
\begin{aligned}
&V_n(\boldsymbol{x}) && \text{(B.1.15)} \\
&\overset{\text{def}}{=} Y_{2,n}(\boldsymbol{x}) - Y_{3,n}(\boldsymbol{x}) \\
&= \frac{1}{\sqrt{h^d f_{\boldsymbol{X}}(\boldsymbol{x}) \sigma_n^2(\boldsymbol{x})}} \int \int_{\Gamma_n} \{\psi(y - \theta_0(\boldsymbol{x})) - \psi(y - \theta_0(\boldsymbol{u}))\} K\left(\frac{\boldsymbol{x} - \boldsymbol{u}}{h}\right) dW\big(T(y, \boldsymbol{u})\big). && \text{(B.1.16)}
\end{aligned}
$$

$\|V_n\| = \mathcal{O}_p\big(h^{1/2-\delta}\big)$ if

$$
\lim_{\eta \to \infty} \mathrm{P}\left\{ \sup_{\boldsymbol{x} \in \mathcal{D}} \left| \frac{V(\boldsymbol{x})}{\sqrt{h}} \right| > \eta h^{-\delta} \right\} = 0, \text{ for all } n \in \mathbb{N}.
$$

Since $\psi(y - \theta_0(\boldsymbol{x})) - \psi(y - \theta_0(\boldsymbol{u})) = \operatorname{sign}(\theta_0(\boldsymbol{u}) - \theta_0(\boldsymbol{x})) \mathbf{1}\big\{[\theta_0(\boldsymbol{x}) \wedge \theta_0(\boldsymbol{u}), \theta_0(\boldsymbol{x}) \vee \theta_0(\boldsymbol{u})]\big\}$, thus

$$
\big\{\psi(y - \theta_0(\boldsymbol{x})) - \psi(y - \theta_0(\boldsymbol{u}))\big\}^2 = \mathbf{1}\big\{[\theta_0(\boldsymbol{x}) \wedge \theta_0(\boldsymbol{u}), \theta_0(\boldsymbol{x}) \vee \theta_0(\boldsymbol{u})]\big\}.
$$

By assumption the conditional distribution function $F_{Y|\boldsymbol{X}}$ and the function θ_0 are both continuously differentiable and change of variables and an application of

the multivariate mean value theorem gives

$$
\begin{aligned}
&\mathsf{E}\left[\left\{\frac{V_n(\boldsymbol{x})}{\sqrt{h}}\right\}^2\right] \\
&= \frac{1}{h^{d+1} f_{\boldsymbol{X}}(\boldsymbol{x})\sigma_n^2(\boldsymbol{x})} \int\int_{\Gamma_n} \{\psi(y-\theta_0(\boldsymbol{x})) - \psi(y-\theta_0(\boldsymbol{u}))\}^2 K^2\left(\frac{\boldsymbol{x}-\boldsymbol{u}}{h}\right) \\
&\qquad\qquad f(y,\boldsymbol{u}) dy d\boldsymbol{u} \\
&\le \frac{1}{h^{d+1} f_{\boldsymbol{X}}(\boldsymbol{x})\sigma_n^2(\boldsymbol{x})} \int \left|F_{Y|\boldsymbol{X}}(\theta_0(\boldsymbol{x})|\boldsymbol{u}) - F_{Y|\boldsymbol{X}}(\theta_0(\boldsymbol{u})|\boldsymbol{u})\right| K^2\left(\frac{\boldsymbol{x}-\boldsymbol{u}}{h}\right) f_{\boldsymbol{X}}(\boldsymbol{u}) d\boldsymbol{u} \\
&= \frac{1}{h f_{\boldsymbol{X}}(\boldsymbol{x})\sigma_n^2(\boldsymbol{x})} \int K^2(\boldsymbol{z}) \left| \sum_{|\boldsymbol{\alpha}|=1} \partial^{\boldsymbol{\alpha}}\Big(F_{Y|\boldsymbol{X}} \circ \theta_0\Big)(\boldsymbol{\xi}) \right| |h\boldsymbol{z}| f_{\boldsymbol{X}}(\boldsymbol{x}) d\boldsymbol{z} + \mathcal{O}(h) \\
&\le \frac{1}{\sigma_n^2(\boldsymbol{x})} \left\| \sum_{|\boldsymbol{\alpha}|=1} \partial^{\boldsymbol{\alpha}}\Big(F_{Y|\boldsymbol{X}} \circ \theta_0\Big) \right\| \left(\int |\boldsymbol{z}| K^2(\boldsymbol{z}) d\boldsymbol{z} \right) + \mathcal{O}(h),
\end{aligned}
$$

where $\boldsymbol{\xi}$ lies on the line connecting $\boldsymbol{x}$ and $\boldsymbol{u}$. Note that $\sigma_n^2(\boldsymbol{x}) \ge \min\{\tau^2, (1-\tau)^2\}$. It follows from the continuous differentiability of $F_{Y|\boldsymbol{X}}$ and θ_0 that $\left\|\partial^{\boldsymbol{\alpha}}(F_{Y|\boldsymbol{X}} \circ \theta_0)\right\|$ is bounded.

$$
\sigma^2 \stackrel{\text{def}}{=} \sup_{\boldsymbol{x}} \mathsf{E}\left[\left(\frac{V_n(\boldsymbol{x})}{\sqrt{h}}\right)^2\right] \le C + \mathcal{O}(h), \tag{B.1.17}
$$

Now we compute $d(\boldsymbol{s},\boldsymbol{t})$ defined in Lemma B.2.3. Again from $\sigma_n^2(\boldsymbol{x}) \ge \min\{\tau^2, (1-\tau)^2\}$ and (A4),

$$
\begin{aligned}
&\mathsf{E}\left[\left(\frac{V_n(\boldsymbol{t}) - V_n(\boldsymbol{s})}{\sqrt{h}}\right)^2\right] \\
&\le C \frac{1}{h^{d+1}} \int\int_{\Gamma_n} \left\{ \big[\psi(y-\theta_0(\boldsymbol{t})) - \psi(y-\theta_0(\boldsymbol{u}))\big] K\left(\frac{\boldsymbol{t}-\boldsymbol{u}}{h}\right) \right. \\
&\quad \left. - \big[\psi(y-\theta_0(\boldsymbol{s})) - \psi(y-\theta_0(\boldsymbol{u}))\big] K\left(\frac{\boldsymbol{s}-\boldsymbol{u}}{h}\right) \right\}^2 f(y,\boldsymbol{u}) dy d\boldsymbol{u} \\
&= C \frac{1}{h^{d+1}} \int\int_{\Gamma_n} \left\{ \big[\psi(y-\theta_0(\boldsymbol{t})) - \psi(y-\theta_0(\boldsymbol{u}))\big] \left[K\left(\frac{\boldsymbol{t}-\boldsymbol{u}}{h}\right) - K\left(\frac{\boldsymbol{s}-\boldsymbol{u}}{h}\right) \right] \right. \\
&\quad \left. - \big[(\psi(y-\theta_0(\boldsymbol{s})) - \psi(y-\theta_0(\boldsymbol{u}))) - (\psi(y-\theta_0(\boldsymbol{t})) - \psi(y-\theta_0(\boldsymbol{u})))\big] K\left(\frac{\boldsymbol{s}-\boldsymbol{u}}{h}\right) \right\}^2 \\
&\qquad\qquad f(y,\boldsymbol{u}) dy d\boldsymbol{u},
\end{aligned}
$$

which implies

$$
\begin{aligned}
&\mathsf{E}\left[\Big(\frac{V_n(\boldsymbol{t})-V_n(\boldsymbol{s})}{\sqrt{h}}\Big)^2\right]\\
&\leq \frac{2C}{h^{d+1}}\Bigg(\int\int_{\Gamma_n}\big[\psi(y-\theta_0(\boldsymbol{t}))-\psi(y-\theta_0(\boldsymbol{s}))\big]^2K^2\left(\frac{\boldsymbol{s}-\boldsymbol{u}}{h}\right)f(y,\boldsymbol{u})dyd\boldsymbol{u}+\\
&\int\int_{\Gamma_n}\big[\psi(y-\theta_0(\boldsymbol{t}))-\psi(y-\theta_0(\boldsymbol{u}))\big]^2\left[K\left(\frac{\boldsymbol{t}-\boldsymbol{u}}{h}\right)-K\left(\frac{\boldsymbol{s}-\boldsymbol{u}}{h}\right)\right]^2f(y,\boldsymbol{u})dyd\boldsymbol{u}\Bigg)\\
&\stackrel{\text{def}}{=} I_1+I_2.
\end{aligned}
$$

Furthermore,

$$
\begin{aligned}
I_1 &\leq \frac{2C}{h^{d+1}}\int\left|F_{Y|\boldsymbol{X}}(\theta_0(\boldsymbol{t})|\boldsymbol{u})-F_{Y|\boldsymbol{X}}(\theta_0(\boldsymbol{s})|\boldsymbol{u})\right|K^2\left(\frac{\boldsymbol{s}-\boldsymbol{u}}{h}\right)f_{\boldsymbol{X}}(\boldsymbol{u})d\boldsymbol{u}\\
&\leq \frac{2CD}{h^{d+1}}\|\boldsymbol{s}-\boldsymbol{t}\|_\infty\int K^2\left(\frac{\boldsymbol{s}-\boldsymbol{u}}{h}\right)f_{\boldsymbol{X}}(\boldsymbol{u})d\boldsymbol{u}\leq\frac{2C'D}{h}\|\boldsymbol{s}-\boldsymbol{t}\|_\infty,
\end{aligned}
$$

where $\|\boldsymbol{s}-\boldsymbol{t}\|_\infty=\sup_j|s_j-t_j|$. A change of variables and the fact that K is bounded yield

$$
\begin{aligned}
&I_2\\
&\leq \frac{2C}{h^{d+1}}\int\left|F_{Y|\boldsymbol{X}}(\theta_0(\boldsymbol{t})|\boldsymbol{u})-F_{Y|\boldsymbol{X}}(\theta_0(\boldsymbol{u})|\boldsymbol{u})\right|\left[K\left(\frac{\boldsymbol{t}-\boldsymbol{u}}{h}\right)-K\left(\frac{\boldsymbol{s}-\boldsymbol{u}}{h}\right)\right]^2f_{\boldsymbol{X}}(\boldsymbol{u})d\boldsymbol{u}\\
&\leq \frac{4C}{h}\frac{\|\boldsymbol{s}-\boldsymbol{t}\|_\infty}{h}\int\left|K(\boldsymbol{z})-K\left(\boldsymbol{z}+\frac{\boldsymbol{s}-\boldsymbol{t}}{h}\right)\right|d\boldsymbol{z}\\
&\leq 4C\frac{\|\boldsymbol{s}-\boldsymbol{t}\|_\infty}{h^2}\left[\int_{[-A,A]^d}|K(\boldsymbol{z})|\,d\boldsymbol{z}+\int_{[-A,A]^d-\frac{\boldsymbol{s}-\boldsymbol{t}}{h}}\left|K\left(\boldsymbol{z}+\frac{\boldsymbol{s}-\boldsymbol{t}}{h}\right)\right|d\boldsymbol{z}\right]\\
&= 4C'\frac{\|\boldsymbol{s}-\boldsymbol{t}\|_\infty}{h^2}.
\end{aligned}
$$

Thus, for the function γ defined in Lemma B.2.3 we obtain the estimate $\gamma(\epsilon)\leq C(\sqrt{\epsilon}/h)$ and thus

$$
Q(m)\leq(2+\sqrt{2})\frac{C}{h}\int_1^\infty\sqrt{m2^{-y^2}}dy\leq C'\frac{\sqrt{m}}{h},
$$

where $C'>0$. Observe that the graph of the inverse of a univariate, injective function $Q(m)$ is its reflection about the line $y=x$, so the inverse of an upper bound for Q would be a lower bound for Q^{-1}. Given the upper bound above, we can therefore bound Q^{-1} from below by

$$
Q^{-1}(a)\geq(C')^{-2}h^2a^2.
$$

We have $Q^{-1}\big(1/(\eta h^{-\delta})\big) \geq (C')^{-2}\eta^{-1}h^{2+2\delta}$. Applying Lemma B.2.3 yields

$$\mathrm{P}\left\{\sup_{\boldsymbol{x}\in\mathcal{D}}\left|\frac{V_n(\boldsymbol{x})}{\sqrt{h_n}}\right| > \eta h_n^{-\delta}\right\} \leq C''\eta^d h_n^{-2d(1+\delta)}\exp\left\{-C'''\eta^2 h_n^{-2\delta}\right\} \to 0,$$

as $\eta \to \infty$ for all $n \in \mathbb{N}$. □

LEMMA B.1.8. $Y_{3,n} \stackrel{\mathcal{L}}{=} Y_{4,n}$.

PROOF. Since both processes are Gaussian with mean zero, we only need to check the equality of the covariance functions of the two processes at any given time points $\boldsymbol{s}, \boldsymbol{t} \in \mathcal{D}$. Ignoring the normalizing factors in the front, the covariance of $Y_{3,n}$ function is:

$$\begin{aligned} r_3(\boldsymbol{s},\boldsymbol{t}) &= \int\int_{\Gamma_n} \psi^2\big(y-\theta_0(\boldsymbol{u})\big)K\left(\frac{\boldsymbol{s}-\boldsymbol{u}}{h}\right)K\left(\frac{\boldsymbol{t}-\boldsymbol{u}}{h}\right)f(y,\boldsymbol{u})dyd\boldsymbol{u} \\ &= \int \mathsf{E}\Big[\psi^2\big(Y_i-\theta_0(\boldsymbol{u})\big)\mathbf{1}(|Y_i|\leq a_n)|\boldsymbol{u}\Big]K\left(\frac{\boldsymbol{s}-\boldsymbol{u}}{h}\right)K\left(\frac{\boldsymbol{t}-\boldsymbol{u}}{h}\right)f_{\boldsymbol{X}}(\boldsymbol{u})d\boldsymbol{u} \\ &= \int \sigma_n^2(\boldsymbol{u})f_{\boldsymbol{X}}(\boldsymbol{u})K\left(\frac{\boldsymbol{s}-\boldsymbol{u}}{h}\right)K\left(\frac{\boldsymbol{t}-\boldsymbol{u}}{h}\right)d\boldsymbol{u} = r_4(\boldsymbol{s},\boldsymbol{t}) \end{aligned}$$

which is, up to a factor, the covariance function of $Y_{4,n}$. □

LEMMA B.1.9. $\|Y_{4,n} - Y_{5,n}\| = \mathcal{O}_p(h^{1-\delta})$, for $0 < \delta < 1$.

PROOF. We will proceed as in Lemma B.1.6 and apply Lemma B.2.3. Set

$$\begin{aligned} &\tilde{Y}(\boldsymbol{x}) \stackrel{\text{def}}{=} Y_{4,n} - Y_{5,n} \\ &= \frac{1}{\sqrt{h^d f_{\boldsymbol{X}}(\boldsymbol{x})\sigma_n^2(\boldsymbol{x})}}\int\left(\sqrt{\sigma_n^2(\boldsymbol{u})f_{\boldsymbol{X}}(\boldsymbol{u})} - \sqrt{\sigma_n^2(\boldsymbol{x})f_{\boldsymbol{X}}(\boldsymbol{x})}\right)K\left(\frac{\boldsymbol{x}-\boldsymbol{u}}{h}\right)dW(\boldsymbol{u}). \end{aligned}$$

Notice that

$$\sigma_n^2(\boldsymbol{u}) = \tau(1-\tau) - \int_{\{|y|>a_n\}}\psi^2\big(y-\theta_0(\boldsymbol{u})\big)f_{Y|\boldsymbol{X}}(y|\boldsymbol{u})dy,$$

where

$$\int_{\{|y|>a_n\}}\psi^2\big(y-\theta_0(\boldsymbol{u})\big)f_{Y|\boldsymbol{X}}(y|\boldsymbol{u})dy \leq \int_{\{|y|>a_n\}} f_{Y|\boldsymbol{X}}(y|\boldsymbol{u})dy.$$

(B.1.1) suggests that

$$\int_{\{|y|>a_n\}} f_{Y|\boldsymbol{X}}(y|\boldsymbol{u})dy = \mathcal{O}(h^d(\log n)^{-1}).$$

Hence, we have $\sigma_n^2(\boldsymbol{u}) \le C_\tau + E_n$, where $E_n = \mathcal{O}(h^d(\log n)^{-1})$, and $C_\tau = \tau(1-\tau)$.

$$
\begin{aligned}
&\mathsf{E}\Big[\Big(\frac{\tilde{Y}(\boldsymbol{t})}{h}\Big)^2\Big] \\
&= \frac{1}{h^{d+2} f_{\boldsymbol{X}}(\boldsymbol{t})\sigma_n^2(\boldsymbol{t})} \int \Big(\sqrt{\sigma_n^2(\boldsymbol{u}) f_{\boldsymbol{X}}(\boldsymbol{u})} - \sqrt{\sigma_n^2(\boldsymbol{t}) f_{\boldsymbol{X}}(\boldsymbol{t})}\Big)^2 K^2\left(\frac{\boldsymbol{t}-\boldsymbol{u}}{h}\right) d\boldsymbol{u} \\
&= \frac{1}{h^{d+2} f_{\boldsymbol{X}}(\boldsymbol{t})\sigma_n^2(\boldsymbol{t})} \int \Big\{\sqrt{\sigma_n^2(\boldsymbol{u})}\Big[\sqrt{f_{\boldsymbol{X}}(\boldsymbol{u})} - \sqrt{f_{\boldsymbol{X}}(\boldsymbol{t})}\Big] \\
&\qquad\qquad + \sqrt{f_{\boldsymbol{X}}(\boldsymbol{x})}\Big[\sqrt{\sigma_n^2(\boldsymbol{u})} - \sqrt{\sigma_n^2(\boldsymbol{t})}\Big]\Big\}^2 K^2\left(\frac{\boldsymbol{t}-\boldsymbol{u}}{h}\right) d\boldsymbol{u} \\
&\le 2Ch^{-d-2}\left\{\max\{\tau^2,(1-\tau)^2\}\int \Big[\sqrt{f_{\boldsymbol{X}}(\boldsymbol{u})} - \sqrt{f_{\boldsymbol{X}}(\boldsymbol{t})}\Big]^2 K^2\left(\frac{\boldsymbol{t}-\boldsymbol{u}}{h}\right) d\boldsymbol{u}\right. \\
&\qquad\qquad \left. + C\int \Big[\sqrt{\sigma_n^2(\boldsymbol{u})} - \sqrt{\sigma_n^2(\boldsymbol{t})}\Big]^2 K^2\left(\frac{\boldsymbol{t}-\boldsymbol{u}}{h}\right) d\boldsymbol{u}\right\},
\end{aligned}
$$

Since

$$
\Big[\sqrt{\sigma_n^2(\boldsymbol{u})} - \sqrt{\sigma_n^2(\boldsymbol{t})}\Big]^2 = \left[\frac{\sigma_n^2(\boldsymbol{u}) - \sigma_n^2(\boldsymbol{t})}{\sqrt{\sigma_n^2(\boldsymbol{u})} + \sqrt{\sigma_n^2(\boldsymbol{t})}}\right]^2 \le CE_n^2 = \mathcal{O}(h^{2d}(\log n)^{-2});
$$

moreover, $\sqrt{f_{\boldsymbol{X}}(\boldsymbol{x})}$ is continuously differentiable on $\mathcal{D}$ by assumption (A4). Along with $\int |z|^2 K(z) < \infty$, we may bound

$$
\sup_{\boldsymbol{t}\in\mathcal{D}} \mathsf{E}\Big[\Big(\frac{\tilde{Y}(\boldsymbol{t})}{h}\Big)^2\Big] \le C + \mathcal{O}(h^{2d-2}(\log n)^{-2}).
$$

On the other hand,

$$
\begin{aligned}
&\mathsf{E}\Big[\Big(\frac{\tilde{Y}(\boldsymbol{t}) - \tilde{Y}(\boldsymbol{s})}{h}\Big)^2\Big] \\
&\le Ch^{-d-2}\int \left\{\Big[\sqrt{\sigma_n^2(\boldsymbol{u}) f_{\boldsymbol{X}}(\boldsymbol{u})} - \sqrt{\sigma_n^2(\boldsymbol{t}) f_{\boldsymbol{X}}(\boldsymbol{t})}\Big] K\left(\frac{\boldsymbol{t}-\boldsymbol{u}}{h}\right)\right. \\
&\qquad\qquad \left. - \Big[\sqrt{\sigma_n^2(\boldsymbol{u}) f_{\boldsymbol{X}}(\boldsymbol{u})} - \sqrt{\sigma_n^2(\boldsymbol{s}) f_{\boldsymbol{X}}(\boldsymbol{s})}\Big] K\left(\frac{\boldsymbol{s}-\boldsymbol{u}}{h}\right)\right\}^2 d\boldsymbol{u} \\
&= Ch^{-d-2}\int \left\{\Big[\sqrt{\sigma_n^2(\boldsymbol{u}) f_{\boldsymbol{X}}(\boldsymbol{u})} - \sqrt{\sigma_n^2(\boldsymbol{t}) f_{\boldsymbol{X}}(\boldsymbol{t})}\Big]\Big[K\left(\frac{\boldsymbol{t}-\boldsymbol{u}}{h}\right) - K\left(\frac{\boldsymbol{s}-\boldsymbol{u}}{h}\right)\Big]\right. \\
&\qquad\qquad \left. + \Big[\sqrt{\sigma_n^2(\boldsymbol{t}) f_{\boldsymbol{X}}(\boldsymbol{t})} - \sqrt{\sigma_n^2(\boldsymbol{s}) f_{\boldsymbol{X}}(\boldsymbol{s})}\Big] K\left(\frac{\boldsymbol{s}-\boldsymbol{u}}{h}\right)\right\}^2 d\boldsymbol{u} \\
&\le 2Ch^{-d-2}\int \Big[\sqrt{\sigma_n^2(\boldsymbol{u}) f_{\boldsymbol{X}}(\boldsymbol{u})} - \sqrt{\sigma_n^2(\boldsymbol{t}) f_{\boldsymbol{X}}(\boldsymbol{t})}\Big]^2 \Big[K\left(\frac{\boldsymbol{t}-\boldsymbol{u}}{h}\right) - K\left(\frac{\boldsymbol{s}-\boldsymbol{u}}{h}\right)\Big]^2 d\boldsymbol{u} \\
&\qquad\qquad + 2Ch^{-d-2}\int \Big[\sqrt{\sigma_n^2(\boldsymbol{t}) f_{\boldsymbol{X}}(\boldsymbol{t})} - \sqrt{\sigma_n^2(\boldsymbol{s}) f_{\boldsymbol{X}}(\boldsymbol{s})}\Big]^2 K^2\left(\frac{\boldsymbol{s}-\boldsymbol{u}}{h}\right) d\boldsymbol{u} \\
&\stackrel{\text{def}}{=} I_1 + I_2.
\end{aligned}
$$

From

$$\Big[\sqrt{\sigma_n^2(\boldsymbol{t})f_{\boldsymbol{X}}(\boldsymbol{t})} - \sqrt{\sigma_n^2(\boldsymbol{s})f_{\boldsymbol{X}}(\boldsymbol{s})}\Big]^2 = \left[\frac{\sigma_n^2(\boldsymbol{t})f_{\boldsymbol{X}}(\boldsymbol{t}) - \sigma_n^2(\boldsymbol{s})f_{\boldsymbol{X}}(\boldsymbol{s})}{\sqrt{\sigma_n^2(\boldsymbol{t})f_{\boldsymbol{X}}(\boldsymbol{t})} + \sqrt{\sigma_n^2(\boldsymbol{s})f_{\boldsymbol{X}}(\boldsymbol{s})}}\right]^2 \leq C\|\boldsymbol{t}-\boldsymbol{s}\|_\infty^2,$$

we obtain

$$I_2 = C\frac{\|\boldsymbol{t}-\boldsymbol{s}\|_\infty^2}{h^2}.$$

By change of variables and a similar argument as to bound I_2 in the proof of Lemma B.1.6, it follows

$$I_1 \leq C\frac{\|\boldsymbol{s}-\boldsymbol{t}\|_\infty}{h^3}.$$

Hence, under the condition that $\|\boldsymbol{s}-\boldsymbol{t}\|_\infty < 1$ and $h \to 0$, we conclude that

$$\mathsf{E}\left[\Big(\frac{\tilde{Y}(\boldsymbol{t}) - \tilde{Y}(\boldsymbol{s})}{h}\Big)^2\right] \leq C\frac{\|\boldsymbol{s}-\boldsymbol{t}\|_\infty}{h^3}. \tag{B.1.18}$$

With the same notations as in Lemma B.2.3, (B.1.18) implies $\gamma(\epsilon) \leq Ch^{-3/2}\sqrt{\epsilon}$, which gives $Q(m) \leq Ch^{-3/2}\sqrt{m}$. Therefore,

$$Q^{-1}(a) \geq Ch^3a^2, \tag{B.1.19}$$

and

$$Q^{-1}\big((\eta h^{-\delta})^{-1}\big) \geq Ch^3\eta^{-2}h^{2\delta}. \tag{B.1.20}$$

Lemma B.2.3 asserts that

$$\mathrm{P}\left\{\sup_{\boldsymbol{x}\in\mathcal{D}}\Big|\frac{\tilde{Y}(\boldsymbol{x})}{h}\Big| > \eta h^{-\delta}\right\} \leq Ch^{-(3+2\delta)d}\eta^{2d}\exp\left\{-h^{-2\delta}\eta^2\right\} \to 0,$$

as $\eta \to \infty$ and $h \to 0$. □

Finally, an application of Theorem 2 of Rosenblatt (1976) to $Y_{5,n}(\boldsymbol{x})$ concludes the proof of Theorem 2.1.

B.1.2 Proof of Theorem 3.2.4

Now let $\rho_\tau(u) = |\tau - \mathbf{1}(u < 0)|u^2$, be the loss function associated to quantile regression. Then $\psi_\tau(u) = -2\big\{\tau - \mathbf{1}(u<0)\big\}|u|$ and

$$g(\boldsymbol{x}) = \frac{\partial}{\partial t}\mathsf{E}[\varphi(Y-t)|\boldsymbol{X}=\boldsymbol{x}]\Big|_{t=\theta_0(\boldsymbol{x})} = -2\big[F_{Y|\boldsymbol{X}}\big(\theta_0(\boldsymbol{x})|\boldsymbol{x}\big)(2\tau-1) - \tau\big].$$

It is obvious that $g(\boldsymbol{x}) > 0$ for $0 < \tau < 1$, and consequently

$$S_{n,0,0}(\boldsymbol{x}) = -2\big[F_{Y|\boldsymbol{X}}\big(\theta_0(\boldsymbol{x})|\boldsymbol{x}\big)(2\tau-1) - \tau\big]f_{\boldsymbol{X}}(\boldsymbol{x}) + \mathcal{O}(h^s).$$

LEMMA B.1.10. $\|Y_n - Y_{0,n}\| = o_p\big((\log n)^{-1/2}\big)$.

PROOF. We have $\|Y_n - Y_{0,n}\| \le \|Y_n - \hat{Y}_{n,0}\| + \|\hat{Y}_{n,0} - Y_{0,n}\|$, where $\hat{Y}_{n,0}$ is defined as in Lemma B.1.3, with $a_n \asymp (h^{-3d}\log n)^{1/(b_1-2)}$. With such a choice we have

$$h^{-3d}\log n \sup_{\boldsymbol{x}\in\mathcal{D}} \left| \int_{|y|>a_n} y^2 f_{Y|\boldsymbol{X}}(y|\boldsymbol{x})dy \right| = \mathcal{O}(1) \tag{B.1.21}$$

which implies $h^{-3d}\log n \int_{|y|>a_n} y^2 f_Y(y)dy = \mathcal{O}(1)$. It follows that $\|Y_n - \hat{Y}_{n,0}\| = o\big((\log n)^{-1/2}\big)$ via similar arguments as in Lemma B.1.3.

Since

$$\mathsf{E}\big[W_{n,i}^2(\boldsymbol{x})\big] \le (\log n)(nh^d)^{-1} C \int_{|y|>a_n} y^2 f_Y(y)dy,$$

we conclude by Markov's inequality that $|V_n(\boldsymbol{x})| \to 0$ for each $\boldsymbol{x} \in \mathcal{D}$.

As to the tightness, we have

$$\begin{aligned} I_1 &\le 4\tau^2 \int\int \mathbf{1}(y > a_n) \Bigg[\sum_{\boldsymbol{\alpha}\in\{0,1\}^d} (-1)^{d-|\boldsymbol{\alpha}|} \big(y - \theta_0(\boldsymbol{s} + \boldsymbol{\alpha}\odot(\boldsymbol{t}-\boldsymbol{s}))\big) \\ &\qquad\qquad K\left(\frac{\boldsymbol{s} + \boldsymbol{\alpha}\odot(\boldsymbol{t}-\boldsymbol{s}) - \boldsymbol{u}}{h}\right)\Bigg] f(y,\boldsymbol{u})dyd\boldsymbol{u} \\ &\le 8\tau^2 \left\{ \big(h^{-d}C\mu(B)\big)^2 \int_{y>a_n} y^2 f_Y(y)dy + \big(h^{-d}C\mu(B)\big)^2 \int_{y>a_n} f_Y(y)dy \right\} \\ &\le 8\tau^2 \big(h^{-d}C\mu(B)\big)^2 \int_{y>a_n} y^2 f_Y(y)dy. \end{aligned}$$

Hence,

$$\mathsf{E}\big[V(B)^2\big]^{1/2} \le (\log n)^{1/2} h^{-3d/2} C \left(\int_{y>a_n} y^2 f_Y(y)dy \right)^{1/2} \mu(B).$$

The desired result follows by similar arguments as those used to prove Lemma B.1.3. □

LEMMA B.1.11. If $n^{-1/6}h^{-d/2-3d/(b_1-2)} = \mathcal{O}(n^{-\nu})$, $\nu > 0$,

$$\|Y_{0,n} - Y_{1,n}\| = \mathcal{O}_p\big(n^{-1/6}h^{-d/2}(\log n)^{\epsilon+(2d+4)/3}a_n\big)$$

for any $\epsilon > 0$.

PROOF. With similar arguments as in Lemma B.1.4, it holds almost surely that

$$
\begin{aligned}
&h^{d/2}n^{1/6}(\log n)^{-\epsilon-(2d+4)/3}a_n^{-1}|Y_{0,n}-Y_{1,n}| \\
&\leq \mathcal{O}(1)\left|\frac{a_n^{-1}}{\sqrt{f_{\boldsymbol{X}}(\boldsymbol{x})\sigma_n^2(\boldsymbol{x})}}\right|\Big\{\big|(\tau-1)(\theta_0(\boldsymbol{x})+a_n)+\tau(a_n-\theta_0(\boldsymbol{x}))\big|\left|\int_{B_{\boldsymbol{x}}} dK((\boldsymbol{x}-\boldsymbol{u})/h)\right| \\
&+\big|\tau(a_n-\theta_0(\boldsymbol{x}))+(\tau-1)(a_n-\theta_0(\boldsymbol{x}))\big|\Big|K\Big(\frac{\boldsymbol{x}-\cdot_2}{h}\Big)\Big|\big(B_{\boldsymbol{x}}\big) \\
&+\big|(\tau-1)(\theta_0(\boldsymbol{x})+a_n)+\tau(a_n-\theta_0(\boldsymbol{x}))\big|\Big|\sum_{\substack{\alpha_1=1\\ \boldsymbol{\alpha}_2\in\{0,1\}^d\\ \boldsymbol{\alpha}_2\neq\mathbf{1}_2}}\int_{(B_{\boldsymbol{x}})_{\boldsymbol{\alpha}_2}}\partial^{\boldsymbol{\alpha}_2}K\big((\boldsymbol{x}-\cdot_2)/h\big)\Big|(B_{\boldsymbol{x}})_{\mathbf{1}_2-\boldsymbol{\alpha}_2} \\
&+\big|\tau(a_n-\theta_0(\boldsymbol{x}))+(\tau-1)(a_n-\theta_0(\boldsymbol{x}))\big|\Big|\sum_{\substack{\alpha_1=0\\ \boldsymbol{\alpha}_2\in\{0,1\}^d\\ \boldsymbol{\alpha}_2\neq\mathbf{0}_2}}\int_{(B_{\boldsymbol{x}})_{\boldsymbol{\alpha}_2}}\partial^{\boldsymbol{\alpha}_2}K\big((\boldsymbol{x}-\cdot_2)/h\big)\Big|(B_{\boldsymbol{x}})_{\mathbf{1}_2-\boldsymbol{\alpha}_2}\Big\},
\end{aligned}
\tag{B.1.22}
$$

by the assumption on the kernel K, (B.1.22) is almost surely bounded bounded. $h^{d/2}n^{1/6}(\log n)^{-\epsilon-(2d+4)/3}=o(1)$ by the choice of a_n given in Lemma B.1.10. □

LEMMA B.1.12. $\|Y_{1,n}-Y_{2,n}\|=\mathcal{O}_p\big(h^{d/2}\big)$.

PROOF. Since $B_n\big(T(y,\boldsymbol{u})\big)=W_n\big(T(y,\boldsymbol{u})\big)-F(y,\boldsymbol{u})W_n(1,...,1)$, let

$$C_K=\Big|\int K(\boldsymbol{v})\,d\boldsymbol{v}\Big|,$$

we obtain by a change of variables and a first order approximation to $f(y,\boldsymbol{x}-h\boldsymbol{v})$:

$$
\begin{aligned}
&\|Y_{1,n}-Y_{2,n}\| \\
&\leq 2h^{d/2}C_K\left\|\frac{1}{\sqrt{f_{\boldsymbol{X}}(\boldsymbol{x})\sigma_n^2(\boldsymbol{x})}}\int_{\Gamma_n}\big|\varphi(y-\theta_0(\boldsymbol{x}))\big|f(y,\boldsymbol{x})dy+\mathcal{O}(h)\right\||W(1,...,1)| \\
&\leq 2h^{d/2}C_K\left\|\frac{1}{\sqrt{f_{\boldsymbol{X}}(\boldsymbol{x})\sigma_n^2(\boldsymbol{x})}}\max\{\tau,1-\tau\}\big|\mathsf{E}\big[|Y_i|\big|\boldsymbol{x}\big]+\theta_0(\boldsymbol{x})\big|+\mathcal{O}(h)\right\||W(1,...,1)|.
\end{aligned}
$$

Note that $|W(1,...,1)|=\mathcal{O}_p(1)$, Y_i has a finite second moment by assumption and θ_0 is uniformly bounded on $\mathcal{D}$. □

LEMMA B.1.13. $\|Y_{2,n}-Y_{3,n}\|=\mathcal{O}_p\big(h^{1-\delta}\big)$, where $0<\delta<1$.

PROOF. Note that the derivative of expectile loss function is $2\big[\mathbf{1}(u\leq 0)-\tau\big]\big|u\big|$, which is Lipschitz continuous with Lipschitz constant $2\max\{\tau,1-\tau\}$. Define $V(\boldsymbol{x})$

as in Lemma B.1.6,

$$
\begin{aligned}
&\mathsf{E}\left[\Big(\frac{V(\boldsymbol{x})}{h}\Big)^2\right] \\
&\quad= \frac{1}{h^{d+2} f_{\boldsymbol{X}}(\boldsymbol{x})\sigma_n^2(\boldsymbol{x})} \\
&\qquad \int\int_{\Gamma_n} \{\varphi(y-\theta_0(\boldsymbol{x})) - \varphi(y-\theta_0(\boldsymbol{u}))\}^2 K\left(\frac{\boldsymbol{x}-\boldsymbol{u}}{h}\right) f(y,\boldsymbol{u}) dy d\boldsymbol{u} \\
&\quad\le \frac{C_{\theta_0} \max\{\tau, 1-\tau\}^2}{h^{d+2} f_{\boldsymbol{X}}(\boldsymbol{x})\sigma_n^2(\boldsymbol{x})} \\
&\qquad \int \left(F_{Y|\boldsymbol{X}}(a_n|\boldsymbol{u}) - F_{Y|\boldsymbol{X}}(-a_n|\boldsymbol{u})\right) |\boldsymbol{x}-\boldsymbol{u}|^2 K^2\left(\frac{\boldsymbol{x}-\boldsymbol{u}}{h}\right) f_{\boldsymbol{X}}(\boldsymbol{u}) d\boldsymbol{u} \\
&\quad\le \frac{C^2}{h^2 f_{\boldsymbol{X}}(\boldsymbol{x})\sigma_n^2(\boldsymbol{x})} \int K^2(\boldsymbol{z}) |h\boldsymbol{z}|^2 f_{\boldsymbol{X}}(\boldsymbol{x}) d\boldsymbol{z} + O(h) \le \frac{2C^2}{\sigma_n^2(\boldsymbol{x})} \|K\|_2^2 + \mathcal{O}(h),
\end{aligned}
$$

$$
\begin{aligned}
&\mathsf{E}\left[\Big(\frac{V(\boldsymbol{t}) - V(\boldsymbol{s})}{h}\Big)^2\right] \\
&\le \frac{2C}{h^{d+2}} \int\int_{\Gamma_n} \left[\varphi(y-\theta_0(\boldsymbol{t})) - \varphi(y-\theta_0(\boldsymbol{s}))\right]^2 K^2\left(\frac{\boldsymbol{s}-\boldsymbol{u}}{h}\right) dF(y,\boldsymbol{u}) + \\
&\frac{2C}{h^{d+2}} \int\int_{\Gamma_n} \left[\varphi(y-\theta_0(\boldsymbol{t})) - \varphi(y-\theta_0(\boldsymbol{u}))\right]^2 \left[K\left(\frac{\boldsymbol{t}-\boldsymbol{u}}{h}\right) - K\left(\frac{\boldsymbol{s}-\boldsymbol{u}}{h}\right)\right]^2 dF(y,\boldsymbol{u}) \\
&\stackrel{\text{def}}{=} I_1 + I_2,
\end{aligned}
$$

where

$$
\begin{aligned}
I_1 &\le \frac{C}{h^{d+2}} \int \|\boldsymbol{t}-\boldsymbol{s}\|_\infty^2 K^2\left(\frac{\boldsymbol{s}-\boldsymbol{u}}{h}\right) f_{\boldsymbol{X}}(\boldsymbol{u}) d\boldsymbol{u} \\
&\le \frac{C}{h^{d+2}} \|\boldsymbol{s}-\boldsymbol{t}\|_\infty^2 \int K^2\left(\frac{\boldsymbol{s}-\boldsymbol{u}}{h}\right) f_{\boldsymbol{X}}(\boldsymbol{u}) d\boldsymbol{u} \le C\frac{\|\boldsymbol{s}-\boldsymbol{t}\|_\infty^2}{h^2} + \mathcal{O}(1).
\end{aligned}
$$

By a change of variables and a similar argument as used to bound I_2 in Lemma B.1.6, we obtain

$$
I_2 \le C\frac{\|\boldsymbol{s}-\boldsymbol{t}\|_\infty}{h^3}.
$$

for $\|\boldsymbol{s}-\boldsymbol{t}\| < 1$. Following the lines of proof of Lemma B.1.6 or Lemma B.1.9 completes the proof of the lemma. □

LEMMA B.1.14. $Y_{3,n} \stackrel{d}{=} Y_{4,n}$

PROOF. The proof resembles the proof for Lemma B.1.8 and is omitted for brevity. □

LEMMA B.1.15. $\|Y_{4,n} - Y_{5,n}\| = \mathcal{O}_p\big(h^{1-\delta}\big)$, where $0 < \delta < 1$.

PROOF. The proof resembles the proof for Lemma B.1.9 by using (B.1.21). The details are omitted for brevity. □

B.1.3 Proof of Lemma 3.2.8

We first show assertion 1.). Let $\tilde{F}_{\varepsilon|\boldsymbol{X}}(v|\boldsymbol{x})$ be defined as

$$\tilde{F}_{\varepsilon|\boldsymbol{X}}(v|\boldsymbol{x}) = n^{-1}\sum_{i=1}^{n} G\left(\frac{v-\varepsilon_i}{h_0}\right) L_{\bar{h}}(\boldsymbol{x}-\boldsymbol{X}_i)/\hat{f}_{\boldsymbol{X}}(\boldsymbol{x}). \tag{B.1.23}$$

Since $\sup_{\boldsymbol{x}\in\mathcal{D}} |\hat{f}_{\boldsymbol{X}}(\boldsymbol{x}) - f_{\boldsymbol{X}}(\boldsymbol{x})| = \mathcal{O}_p(\bar{h}^s + (n\bar{h}^d)^{-1/2}\log n)$, linearisation yields

$$\tilde{F}_{\varepsilon|\boldsymbol{X}}(v|\boldsymbol{x}) = \frac{\tilde{M}(v,\boldsymbol{x})}{f_{\boldsymbol{X}}(\boldsymbol{x})} + R_n,$$

where $R_n = \mathcal{O}_p(\bar{h}^2 + (n\bar{h}^d)^{-1/2}\log n)$ uniformly over $\boldsymbol{x}\in\mathcal{D}$ by assumption (B2), where $\tilde{M}(v,\boldsymbol{x}) = \tilde{F}_{\varepsilon|\boldsymbol{X}}(v|\boldsymbol{x})\hat{f}_{\boldsymbol{X}}(\boldsymbol{x}) = n^{-1}\sum_{i=1}^{n} G\left(\frac{v-\varepsilon_i}{h_0}\right) L_{\bar{h}}(\boldsymbol{x}-\boldsymbol{X}_i)$. By Theorem 6.2. (i) of Li and Racine (2007), $\mathsf{E}\big[\tilde{M}(v,\boldsymbol{x}) - F_{\varepsilon,\boldsymbol{X}}(v,\boldsymbol{x})\big]$ is of order $\mathcal{O}(h_0^2 + d\bar{h}^2)$. It remains to show that

$$\sup_{v\in I}\sup_{\boldsymbol{x}\in\mathcal{D}} \left|\tilde{M}(v,\boldsymbol{x}) - \mathsf{E}\big[\tilde{M}(v,\boldsymbol{x})\big]\right| = \mathcal{O}_p\left((n\bar{h}^d)^{-1/2}\log n\right). \tag{B.1.24}$$

By Theorem 6.2. (ii) of Li and Racine (2007), $\mathrm{Var}\left(\tilde{M}(v,\boldsymbol{x})\right) = \mathcal{O}\big\{(n\bar{h}^d)^{-1}\big\}$. By virtue of a standard δ_n-net discretization argument and the Bernstein inequality we obtain (B.1.24).

Next we show that $\left|\hat{F}_{\varepsilon|\boldsymbol{X}}(v|\boldsymbol{x}) - \tilde{F}_{\varepsilon|\boldsymbol{X}}(v|\boldsymbol{x})\right| = \mathcal{O}_p(h^2 + (nh^d)^{-1/2}\log n)$. We have

$$\begin{aligned}
&\hat{F}_{\varepsilon|\boldsymbol{X}}(v|\boldsymbol{x}) - \tilde{F}_{\varepsilon|\boldsymbol{X}}(v|\boldsymbol{x}) \\
&= \frac{1}{n\hat{f}_{\boldsymbol{X}}(\boldsymbol{x})}\sum_{i=1}^{n}\left\{G\left(\frac{v-\varepsilon_i}{h_0}\right) - G\left(\frac{v-\hat{\varepsilon}_i}{h_0}\right)\right\} L_{\bar{h}}(\boldsymbol{x}-\boldsymbol{X}_i) \\
&= \frac{1}{n\hat{f}_{\boldsymbol{X}}(\boldsymbol{x})}\sum_{i=1}^{n}\left\{h_0^{-1} g\left(\frac{v-\varepsilon_i}{h_0}\right)(\varepsilon_i - \hat{\varepsilon}_i)\right\} L_{\bar{h}}(\boldsymbol{x}-\boldsymbol{X}_i) + R_{1,n},
\end{aligned}$$

where $R_{1,n}$ is of negligible order by (B1) under the claim in Section 3.3 of Muhsal and Neumeyer (2010). $\varepsilon_i - \hat{\varepsilon}_i = \hat{\theta}_n(\boldsymbol{X}_i) - \theta_0(\boldsymbol{X}_i)$, which is stochastically bounded with $h^s + (nh^d)^{-1/2}\log n$, for arbitrary $\delta > 0$. Moreover, observe that

$$\frac{1}{n}\sum_{i=1}^{n} h_0^{-1} g\left(\frac{v-\varepsilon_i}{h_0}\right) L_{\bar{h}}(\boldsymbol{x}-\boldsymbol{X}_i)$$

is a kernel density estimator which has standard bias and variance and which is is stochastically bounded. Hence, in order to estimate

$$\mathrm{P}\left\{\left|\hat{F}_{\varepsilon|\boldsymbol{X}}(v|\boldsymbol{x}) - \tilde{F}_{\varepsilon|\boldsymbol{X}}(v|\boldsymbol{x})\right| > \eta n^{-\lambda}\right\},$$

splitting the probability of under the event

$$\left\{\left|\hat{\theta}_n(\boldsymbol{X}_i) - \theta_0(\boldsymbol{X}_i)\right| > h^s + (nh^d)^{-1/2}\log n\right\}$$

and its complement, where $n^{-\lambda} = h_0^2 + h^s + \bar{h}^2 + (nh_0\bar{h}^d)^{-1/2}\log n + (nh^d)^{-1/2}\log n$, we get the desired result.

Next we show assertion 2.). Let $\tilde{f}_{\varepsilon|\boldsymbol{X}}(v|\boldsymbol{x})$ be defined as

$$\tilde{f}_{\varepsilon|\boldsymbol{X}}(v|\boldsymbol{x}) = n^{-1}\sum_{i=1}^{n} g_{h_0}(v-\varepsilon_i)\, L_{\bar{h}}(\boldsymbol{x}-\boldsymbol{X}_i)/\hat{f}_{\boldsymbol{X}}(\boldsymbol{x}). \tag{B.1.25}$$

By standard theory for kernel density estimation, we have

$$\tilde{f}_{\varepsilon|\boldsymbol{X}}(v|\boldsymbol{x}) = \frac{\tilde{m}(v,\boldsymbol{x})}{f_{\boldsymbol{X}}(\boldsymbol{x})} + R_n,$$

where $R_n = \mathcal{O}_p(\bar{h}^s + (n\bar{h}^d)^{-1/2}\log n)$ uniformly over $\boldsymbol{x}\in\mathcal{D}$ by assumption (B2), where $\tilde{m}(v,\boldsymbol{x}) = \tilde{f}_{\varepsilon|\boldsymbol{X}}(v|\boldsymbol{x})\hat{f}_{\boldsymbol{X}}(\boldsymbol{x}) = n^{-1}\sum_{i=1}^{n} g_{h_0}(\varepsilon_i - v)\, L_{\bar{h}}(\boldsymbol{x}-\boldsymbol{X}_i)$. It follows from the standard theory of density estimation that

$$\|\tilde{m}(v,\boldsymbol{x}) - f_{\varepsilon,\boldsymbol{X}}(v,\boldsymbol{x})\| = \mathcal{O}_p(h_0^2 + \bar{h}^2 + (nh_0\bar{h}^d)^{-1/2}\log n). \tag{B.1.26}$$

A Taylor expansion yields

$$\begin{aligned}
&\tilde{f}_{\varepsilon|\boldsymbol{X}}(v|\boldsymbol{x}) - \hat{f}_{\varepsilon|\boldsymbol{X}}(v|\boldsymbol{x}) \\
&= \frac{1}{n\hat{f}_{\boldsymbol{X}}(\boldsymbol{x})}\sum_{i=1}^{n}\left\{g_{h_0}(v-\varepsilon_i) - g_{h_0}(v-\hat{\varepsilon}_i)\right\} L_{\bar{h}}(\boldsymbol{x}-\boldsymbol{X}_i) \\
&= \frac{1}{n\hat{f}_{\boldsymbol{X}}(\boldsymbol{x})}\sum_{i=1}^{n}\left\{h_0^{-2}g'\left(\frac{v-\varepsilon_i}{h_0}\right)\left(\hat{\theta}_n(\boldsymbol{X}_i) - \theta_0(\boldsymbol{X}_i)\right)\right\} L_{\bar{h}}(\boldsymbol{x}-\boldsymbol{X}_i) + R_{2,n}
\end{aligned}$$

it follows from Muhsal and Neumeyer (2010) that $R_{2,n}$ is negligible under condition (B1). Again

$$\frac{1}{n}\sum_{i=1}^{n} h_0^{-2}g'\left(\frac{v-\varepsilon_i}{h_0}\right) L_{\bar{h}}(\boldsymbol{x}-\boldsymbol{X}_i)$$

is a kernel estimator for the derivative of the conditional density function and is thus stochastically bounded. Applying the stochastic bound for $\hat{\theta}_n(\boldsymbol{X}_i) - \theta_0(\boldsymbol{X}_i)$ and similar probability separating argument for proving 1.), assertion 2.) follows.

For the third estimator 3.), define

$$\tilde{\sigma}^2(\boldsymbol{x}) = n^{-1}\sum_{i=1}^{n}\psi^2(\varepsilon_i) L_{\bar{h}}(\boldsymbol{x}-\boldsymbol{X}_i)/\hat{f}_{\boldsymbol{X}}(\boldsymbol{x}).$$

Using a weak uniform consistency result for kernel regression, see, for instance, Hansen (2008), $\left\|\tilde{\sigma}^2(\boldsymbol{x}) - \sigma^2(\boldsymbol{x})\right\| = \mathcal{O}_p\big(\bar{h}^2 + (n\bar{h}^d)^{-1/2}\log n\big)$. Below we separately discuss the quantile and expectile case.

In the quantile case, $\psi(u) = \mathbf{1}(u<0) - \tau$, then

$$\begin{aligned}
\hat{\sigma}^2(\boldsymbol{x}) - \tilde{\sigma}^2(\boldsymbol{x}) &= n^{-1}\sum_{i=1}^{n}\left[\psi^2(\hat{\varepsilon}_i) - \psi^2(\varepsilon_i)\right] L_{\bar{h}}(\boldsymbol{x}-\boldsymbol{X}_i) \\
&= (1-2\tau)n^{-1}\sum_{i=1}^{n}\left[\mathbf{1}(\hat{\varepsilon}_i<0) - \mathbf{1}(\varepsilon_i<0)\right] L_{\bar{h}}(\boldsymbol{x}-\boldsymbol{X}_i).
\end{aligned}$$

Note that $\mathbf{1}(\hat{\varepsilon}_i < 0) - \mathbf{1}(\varepsilon_i < 0) = \mathbf{1}(\theta_0(\boldsymbol{X}_i) < Y_i < \hat{\theta}_n(\boldsymbol{X}_i)) - \mathbf{1}(\hat{\theta}_n(\boldsymbol{X}_i) < Y_i < \theta_0(\boldsymbol{X}_i))$. Applying the fact that $\sup_{\boldsymbol{x}\in\mathcal{D}} |\hat{\theta}_n(\boldsymbol{x}) - \theta_0(\boldsymbol{x})|$ stochastically bounded, we first restrict our focus on the event $\hat{\theta}_n(\boldsymbol{X}_i) - \theta_0(\boldsymbol{X}_i) < h^s + (nh^d)^{-1/2}\log n$. If $\tau = 1/2$, then $\psi^2(\hat{\varepsilon}_i) - \psi^2(\varepsilon_i) = 0$ and we are done. Given $\tau \neq 1/2$,

$$\begin{aligned}
&(1-2\tau)^{-1}\mathsf{E}\big[\hat{\sigma}^2(\boldsymbol{x}) - \tilde{\sigma}^2(\boldsymbol{x})\big] \\
&= \mathsf{E}\left[\big(\mathbf{1}(\hat{\varepsilon}_i < 0) - \mathbf{1}(\varepsilon_i < 0)\big) L_{\bar{h}}(\boldsymbol{x} - \boldsymbol{X}_i)\right] \\
&= 2\int \Big\{F(\hat{\theta}_n(\boldsymbol{u})|\boldsymbol{u}) - F(\theta_0(\boldsymbol{u})|\boldsymbol{u})\Big\} L_{\bar{h}}(\boldsymbol{x} - \boldsymbol{u}) f_{\boldsymbol{X}}(\boldsymbol{u}) d\boldsymbol{u} \\
&= 2\int f(\theta^{\dagger}(\boldsymbol{u})|\boldsymbol{u})(\hat{\theta}_n(\boldsymbol{u}) - \theta_0(\boldsymbol{u})) L_{\bar{h}}(\boldsymbol{x} - \boldsymbol{u}) f_{\boldsymbol{X}}(\boldsymbol{u}) d\boldsymbol{u},
\end{aligned}$$

where $\theta^{\dagger}(\boldsymbol{u})$ lies between $\hat{\theta}_n(\boldsymbol{u})$ and $\theta_0(\boldsymbol{u})$. By condition (B2), $f(v|\boldsymbol{x})$ is uniformly bounded, we deduce that $\mathsf{E}\big[\hat{\sigma}^2(\boldsymbol{x}) - \tilde{\sigma}^2(\boldsymbol{x})\big] = \mathcal{O}(h^s + (nh^d)^{-1/2}\log n)$. Observe that $(\mathbf{1}(\hat{\varepsilon}_i < 0) - \mathbf{1}(\varepsilon_i < 0))^2 = \mathbf{1}\big\{[\hat{\theta}_n(\boldsymbol{X}_i) \wedge \theta_0(\boldsymbol{X}_i), \hat{\theta}_n(\boldsymbol{X}_i) \vee \theta_0(\boldsymbol{X}_i)]\big\}$. It follows from similar computations

$$\mathsf{E}\big[\big\{\hat{\sigma}^2(\boldsymbol{x}) - \tilde{\sigma}^2(\boldsymbol{x})\big\}^2\big] = \mathcal{O}\big(h^s + (nh^d)^{-1/2}\log n\big).$$

Again observe that $\mathbf{1}\big\{[\hat{\theta}_n(\boldsymbol{X}_i) \wedge \theta_0(\boldsymbol{X}_i), \hat{\theta}_n(\boldsymbol{X}_i) \vee \theta_0(\boldsymbol{X}_i)]\big\}$ is independent of the variable $\boldsymbol{x}$, a discretization argument and the Bernstein inequality yield the result that $n^{\lambda_1} \cdot \|\hat{\sigma}^2(\boldsymbol{x}) - \tilde{\sigma}^2(\boldsymbol{x})\|$ is stochastically bounded.

For the expectile case, $\psi(u) = 2\big[\mathbf{1}(u < 0) - \tau\big]|u|$. Since

$$\begin{aligned}
\psi^2(\varepsilon_i) - \psi^2(\hat{\varepsilon}_i) &= 4\big\{\mathbf{1}(\varepsilon_i < 0) - \tau\big\}^2|\varepsilon_i|^2 - 4\big\{\mathbf{1}(\hat{\varepsilon}_i < 0) - \tau\big\}^2|\hat{\varepsilon}_i|^2 \\
&= 4\big\{\mathbf{1}(\hat{\varepsilon}_i < 0) - \tau\big\}^2\big(|\varepsilon_i|^2 - |\hat{\varepsilon}_i|^2\big) + 4\big\{\mathbf{1}(\varepsilon_i < 0) - \mathbf{1}(\hat{\varepsilon}_i < 0)\big\}^2|\varepsilon_i|^2,
\end{aligned}$$

Thus,

$$\begin{aligned}
\hat{\sigma}^2(\boldsymbol{x}) - \tilde{\sigma}^2(\boldsymbol{x}) &= 4n^{-1}\sum_{i=1}^{n}\big\{\mathbf{1}(\hat{\varepsilon}_i < 0) - \tau\big\}\big(|\varepsilon_i|^2 - |\hat{\varepsilon}_i|^2\big) L_{\bar{h}}(\boldsymbol{x} - \boldsymbol{X}_i) \\
&\quad + 4(1-2\tau)n^{-1}\sum_{i=1}^{n}\big\{\mathbf{1}(\varepsilon_i < 0) - \mathbf{1}(\hat{\varepsilon}_i < 0)\big\}|\varepsilon_i|^2 L_{\bar{h}}(\boldsymbol{x} - \boldsymbol{X}_i) \\
&\stackrel{\text{def}}{=} 4R_{3,n}(\boldsymbol{x}) + 4(1-2\tau)R_{4,n}(\boldsymbol{x}).
\end{aligned}$$

Again, it is sufficient to focus on the set $\{|\hat{\theta}_n(\boldsymbol{X}_i) - \theta_0(\boldsymbol{X}_i)| < n^{-\lambda_0}\}$, where $n^{-\lambda_0} \sim h^s + (nh^d)^{-1/2}\log n$. For $R_{3,n}(\boldsymbol{x})$, notice that

$$\begin{aligned}
|\varepsilon_i|^2 - |\hat{\varepsilon}_i|^2 &= \big(\theta_0(\boldsymbol{X}_i) - \hat{\theta}_n(\boldsymbol{X}_i)\big)\big(\theta_0(\boldsymbol{X}_i) + \hat{\theta}_n(\boldsymbol{X}_i) - 2Y_i\big) \\
&= R_{5,n}(\boldsymbol{u})\big(2\theta_0(\boldsymbol{u}) + R_{5,n}(\boldsymbol{u}) - 2Y_i\big),
\end{aligned}$$

where $\sup_{\boldsymbol{x}\in\mathcal{D}}|R_{5,n}(\boldsymbol{x})| = \mathcal{O}(n^{-\lambda_0})$, so

$$\begin{aligned}
&\mathsf{E}R_{3,n}(\boldsymbol{x}) \\
&= \mathsf{E}\left[\big\{\mathbf{1}(\hat{\varepsilon}_i < 0) - \tau\big\}\big(\theta_0(\boldsymbol{X}_i) - \hat{\theta}_n(\boldsymbol{X}_i)\big)\big(\theta_0(\boldsymbol{X}_i) + \hat{\theta}_n(\boldsymbol{X}_i) - 2Y_i\big)L_{\bar{h}}(\boldsymbol{x} - \boldsymbol{X}_i)\right] \\
&= (1-\tau)^2 \\
&\quad \int\int_{y<\hat{\theta}_n(\boldsymbol{u})} R_{5,n}(\boldsymbol{u})(2\theta_0(\boldsymbol{u}) + R_{5,n}(\boldsymbol{u}) - 2y)L_{\bar{h}}(\boldsymbol{x}-\boldsymbol{u})f_{Y|\boldsymbol{X}}(y|\boldsymbol{u})f_{\boldsymbol{X}}(\boldsymbol{u})dyd\boldsymbol{u} \\
&- \tau^2\int\int_{y>\hat{\theta}_n(\boldsymbol{u})} R_{5,n}(\boldsymbol{u})(2\theta_0(\boldsymbol{u}) + R_{5,n}(\boldsymbol{u}) - 2y)L_{\bar{h}}(\boldsymbol{x}-\boldsymbol{u})f_{Y|\boldsymbol{X}}(y|\boldsymbol{u})f_{\boldsymbol{X}}(\boldsymbol{u})dyd\boldsymbol{u}.
\end{aligned}$$

Hence, $|\mathsf{E}R_{3,n}(\boldsymbol{x})| < Cn^{-\lambda_0}$ for some constant C.

$$\begin{aligned}
\mathrm{Var}\left\{R_{3,n}(\boldsymbol{x})\right\} &\le n^{-1}\max\{(1-\tau)^2, \tau^2\}^2 \\
&\quad \int\int R_{5,n}^2(\boldsymbol{u})(2\theta_0(\boldsymbol{u}) + R_{5,n}(\boldsymbol{u}) - 2y)^2 L_{\bar{h}}^2(\boldsymbol{x}-\boldsymbol{u})f_{Y|\boldsymbol{X}}(y|\boldsymbol{u})f_{\boldsymbol{X}}(\boldsymbol{u})dyd\boldsymbol{u} \\
&\le C(n\bar{h}^d)^{-1}n^{-2\lambda_0}.
\end{aligned}$$

One can apply discretization and the Bernstein inequality to show that

$$\sup_{\boldsymbol{x}\in\mathcal{D}}\left|R_{3,n}(\boldsymbol{x})\right| = \mathcal{O}_p(n^{-\lambda_0}\log n).$$

For $R_{4,n}(\boldsymbol{x})$, again suppose without loss of generality that $\{|\hat{\theta}_n(\boldsymbol{X}_i) - \theta_0(\boldsymbol{X}_i)| < n^{-\lambda_0}\}$, where $n^{-\lambda_0} \sim h^s + (nh^d)^{-1/2}\log n$,

$$\begin{aligned}
|\mathsf{E}[R_{4,n}(\boldsymbol{x})]| &\le 2\int\left[\int_{|y-\theta_0(\boldsymbol{u})|<R_{6,n}(\boldsymbol{u})} |y-\theta_0(\boldsymbol{u})|^2 f_{Y|\boldsymbol{X}}(y|\boldsymbol{u})dy\right]|L_{\bar{h}}(\boldsymbol{x}-\boldsymbol{u})|f_{\boldsymbol{X}}(\boldsymbol{u})d\boldsymbol{u} \\
&= \mathcal{O}(n^{-2\lambda_0}).
\end{aligned}$$

An application of Markov's inequality yields the desired result.

B.1.4 Proof of Theorem 3.3.1

Proof of Lemma 3.3.2. We will discuss the case of quantile and expectile regression separately.

Consider first $\psi(u) = \mathbf{1}(u < 0) - \tau$.

$$\sigma_*^2(\boldsymbol{x}) - \hat{\sigma}^2(\boldsymbol{x}) = n^{-1}\sum_{i=1}^n\left\{\int\psi^2(v)g_{h_0}(v-\hat{\varepsilon}_i) - \psi^2(\hat{\varepsilon}_i)\right\}L_{\bar{h}}(\boldsymbol{x}-\boldsymbol{X}_i)/\hat{f}_{\boldsymbol{X}}(\boldsymbol{x}). \tag{B.1.27}$$

By a change of variables,

$$\begin{aligned}
&\left|\int \psi^2(v) g_{h_0}(v-\hat{\varepsilon}_i)dv - \psi^2(\hat{\varepsilon}_i)\right| \le \int \left|\psi^2(\hat{\varepsilon}_i + wh_0) - \psi^2(\hat{\varepsilon}_i)\right| g(w)dw \\
&\le 2\max\{\tau, 1-\tau\} \int \left|\psi(\hat{\varepsilon}_i + wh_0) - \psi(\hat{\varepsilon}_i)\right| g(w)dw \\
&= C_\tau \Bigg\{ \mathbf{1}(\hat{\varepsilon}_i > \log(n)\cdot h_0) \int_{-\infty}^{-\hat{\varepsilon}_i/h_0} g(w)dw + \mathbf{1}(\hat{\varepsilon}_i < -\log(n)\cdot h_0) \int_{-\hat{\varepsilon}_i/h_0}^{\infty} g(w)dw \\
&+ \mathbf{1}(|\hat{\varepsilon}_i| \le \log(n)\cdot h_0) \int_{\mathbb{R}} g(w)dw \Bigg\} \\
&\le C_\tau \Bigg\{ \mathbf{1}(\hat{\varepsilon}_i > \log(n)\cdot h_0) \int_{-\infty}^{-\log(n)} g(w)dw + \mathbf{1}(\hat{\varepsilon}_i < -\log(n)\cdot h_0) \int_{\log(n)}^{\infty} g(w)dw \\
&+ \mathbf{1}(|\hat{\varepsilon}_i| \le \log(n)\cdot h_0) \Bigg\} \\
&\le C_\tau \Bigg\{ \int_{-\infty}^{-\log(n)} g(w)dw + \int_{\log(n)}^{\infty} g(w)dw + \mathbf{1}(|\hat{\varepsilon}_i| \le \log(n)\cdot h_0) \Bigg\}.
\end{aligned}$$

Hence, the sup norm of (B.1.27) is bounded by $I_1 + I_2 + \sup_{\boldsymbol{x}} |I_3(\boldsymbol{x})|$, where $I_1 \stackrel{\text{def}}{=} C_\tau G(-\log n)$, $I_2 \stackrel{\text{def}}{=} C_\tau\big(1 - G(\log n)\big)$ and

$$I_3(\boldsymbol{x}) \stackrel{\text{def}}{=} n^{-1} \sum_{i=1}^{n} \mathbf{1}(|\hat{\varepsilon}_i| \le h_0 \log n) \big| L_{\bar{h}}(\boldsymbol{x} - \boldsymbol{X}_i)\big| / |\hat{f}_{\boldsymbol{X}}(\boldsymbol{x})|,$$

since $\hat{f}_{\boldsymbol{X}}(\boldsymbol{x}) = n^{-1}\sum_{i=1}^{n} L_{\bar{h}}(\boldsymbol{x} - \boldsymbol{X}_i)$. I_1 and I_2 decay polynomially in n by assumption (A1). Note that for any $\kappa > 0$,

$$\begin{aligned}
&\mathrm{P}\left\{ \sup_{\boldsymbol{x}} \left| (1-\mathsf{E}) \sum_{i=1}^{n} \mathbf{1}(|\hat{\varepsilon}_i| \le h_0 \log n) |L_{\bar{h}}(\boldsymbol{x} - \boldsymbol{X}_i)| \right| > n(\log n)^{-1}\kappa \right\} \le \\
&\mathrm{P}\Bigg\{ \sup_{\boldsymbol{x}} \left| (1-\mathsf{E}) \sum_{i=1}^{n} \mathbf{1}(|\hat{\varepsilon}_i| \le h_0 \log n) |L_{\bar{h}}(\boldsymbol{x} - \boldsymbol{X}_i)| \right| > n(\log n)^{-1}\kappa, \\
&\qquad\qquad \big\|\hat{\theta}_n(\boldsymbol{x}) - \theta_0(\boldsymbol{x})\big\| \le E_n \log n \Bigg\} \\
&\qquad + \mathrm{P}\left\{ \big\|\hat{\theta}_n(\boldsymbol{x}) - \theta_0(\boldsymbol{x})\big\| > E_n \log n \right\}, \qquad \text{(B.1.28)}
\end{aligned}$$

where $E_n = h^s + (nh^d)^{-1/2}\log n$. The uniform convergence of $\hat{\theta}_n(\boldsymbol{x})$ to $\theta_0(\boldsymbol{x})$ yields that

$$\sum_{n=1}^{\infty} \mathrm{P}\left\{ \big\|\hat{\theta}_n(\boldsymbol{x}) - \theta_0(\boldsymbol{x})\big\| > E_n \log n \right\} < \infty. \qquad \text{(B.1.29)}$$

For the first probability, it is easy to see that it is bounded by the sum

$$
\begin{aligned}
&A_n + B_n \\
&\stackrel{\text{def}}{=} \mathrm{P}\left\{ \sup_{\boldsymbol{x}} \left| (1-\mathsf{E}) \sum_{i=1}^{n} \mathbf{1}(|\varepsilon_i| \leq h_0 \log n + E_n \log n) | L_{\bar{h}}(\boldsymbol{x} - \boldsymbol{X}_i) \right| > \frac{1}{2} n (\log n)^{-1} \kappa \right\} \\
&+ \mathbf{1}\Bigg(\sup_{\boldsymbol{x}} \left| \mathsf{E}\left[\sum_{i=1}^{n} \mathbf{1}(h_0 \log n - E_n \log n < |\varepsilon_i| \leq h_0 \log n + E_n \log n) | L_{\bar{h}}(\boldsymbol{x} - \boldsymbol{X}_i) \right] \right| \\
&\qquad\qquad > \frac{1}{2} n (\log n)^{-1} \kappa \Bigg).
\end{aligned}
$$

After an explicit computation of the expectation, one concludes that B_n is equal to zero for any $\kappa > 0$ if n is sufficiently large. Now we need to bound A_n. Note that for any fixed $\boldsymbol{x}$, we can estimate the variance by

$$
\mathrm{Var}\left(\sum_{i=1}^{n} \mathbf{1}(|\varepsilon_i| \leq h_0 \log n + E_n \log n) | L_{\bar{h}}(\boldsymbol{x} - \boldsymbol{X}_i) | \right) \leq C_L n h_0 \bar{h}^{-d} \log n,
$$

applying a concentration inequality, one gets for any $\kappa > 0$,

$$
\begin{aligned}
&\mathrm{P}\left\{ (1-\mathsf{E}) \sum_{i=1}^{n} \mathbf{1}(|\varepsilon_i| \leq h_0 \log n + E_n \log n) | L_{\bar{h}}(\boldsymbol{x} - \boldsymbol{X}_i) | > n (\log n)^{-1} \kappa \right\} \\
&\qquad\qquad \leq 2 \exp\left\{ -\frac{1}{4} \frac{n^2 (\log n)^{-4} \kappa^2}{C_L n h_0 \bar{h}^{-d} \log n + C_L n \bar{h}^{-d} (\log n)^{-2} \kappa} \right\},
\end{aligned}
$$

which decreases exponentially in n since $n\bar{h}^d \to \infty$ polynomially in n by assumption (B3). By a discretization argument, one can show that A_n is also summable (the grid size grows polynomially in n). Hence, we conclude that the probability (B.1.28) is summable. The stochastic part of the numerator of $I_3(\boldsymbol{x})$ is therefore of $\mathcal{O}_p((\log n)^{-1})$ a.s. by an application of the Borel-Cantelli lemma.

The mean of the numerator of $I_3(\boldsymbol{x})$ can be estimated by the law of iterative expectation:

$$
\begin{aligned}
&\mathsf{E}\left[\mathsf{E}\left[n^{-1} \sum_{i=1}^{n} \mathbf{1}(|\hat{\varepsilon}_i| \leq h_0 \log n) \big| L_{\bar{h}}(\boldsymbol{x} - \boldsymbol{X}_i) \big| \,\middle|\, X, \hat{\theta}_n(\boldsymbol{x}) - \theta_0(\boldsymbol{x}) \right] \right] \\
&= \mathsf{E}\left[\int_{\hat{\theta}_n(\boldsymbol{x}) - \theta_0(\boldsymbol{x}) - h_0 \log n}^{\hat{\theta}_n(\boldsymbol{x}) - \theta_0(\boldsymbol{x}) + h_0 \log n} f\big\{ e | X, \hat{\theta}_n(\boldsymbol{x}) - \theta_0(\boldsymbol{x}) \big\} de | L_{\bar{h}}(\boldsymbol{x} - \boldsymbol{X}_i) | \right] \\
&\leq 2 h_0 \log n C = o((\log n)^{-1}),
\end{aligned}
$$

since the density $f\big\{ e | X, \hat{\theta}_n(\boldsymbol{x}) - \theta_0(\boldsymbol{x}) \big\}$ is bounded and $L \in L^1(\mathbb{R}^d)$. Finally, applying a linearization argument we obtain that $\| I_3(\boldsymbol{x}) \| = \mathcal{O}_p((\log n)^{-1}) = o((\log n)^{-1/2})$ a.s.

In the case of expectile regression, we need to consider $\psi(u) = 2(\mathbf{1}(u<0)-\tau)|u|$, which is Lipschitz continuous (see Lemma B.1.13). Note that $|\hat{\varepsilon}_i| \leq |\varepsilon_i| + E_n$, where $E_n = \mathcal{O}(h^s + (nh^d)^{-1/2}\log n)$ a.s. by the Bahadur representation of θ_n, a discretization argument and an application of the Bernstein inequality. Hence,

$$\begin{aligned}
&\left| n^{-1}\sum_{i=1}^{n}\left\{\int \psi^2(v)g_{h_0}(v-\hat{\varepsilon}_i) - \psi^2(\hat{\varepsilon}_i)dv\right\} L_{\bar{h}}(\boldsymbol{x}-\boldsymbol{X}_i)\right| \\
&\leq n^{-1}C_\tau \sum_{i=1}^{n}\int h_0(2|\hat{\varepsilon}_i| + h_0|w|)|w|g(w)dw|L_{\bar{h}}(\boldsymbol{x}-\boldsymbol{X}_i)| \\
&= C_\tau h_0^2 \int |w|^2 g(w)dw + C_{\tau,g}2h_0 n^{-1}\sum_{i=1}^{n}|\hat{\varepsilon}_i||L_{\bar{h}}(\boldsymbol{x}-\boldsymbol{X}_i)| \\
&\leq C_\tau h_0^2 \int |w|^2 g(w)dw + 2C_{\tau,g}h_0 E_n n^{-1}\sum_{i=1}^{n}|L_{\bar{h}}(\boldsymbol{x}-\boldsymbol{X}_i)| \\
&\qquad\qquad + 2C_{\tau,g}h_0 n^{-1}\sum_{i=1}^{n}|\varepsilon_i||L_{\bar{h}}(\boldsymbol{x}-\boldsymbol{X}_i)|.
\end{aligned}$$

The first term converges almost surely to 0, faster than $(\log n)^{-1}$, based on assumption (B3). The second term and the third term can be handled by similar argument for showing the uniform almost sure convergence of the Nadaraya-Watson estimator, see Hansen (2008) for more details.

□

Our strategy is to follow the sequence of approximation steps that are similar to Section B.1.1 and B.1.2. Define

$$Y_{0,n}^*(\boldsymbol{x}) = \frac{1}{\sqrt{h^d \hat{f}_{\boldsymbol{X}}(\boldsymbol{x})\sigma_{n,*}^2(\boldsymbol{x})}}\int\int_{\Gamma_n^*} K\left(\frac{\boldsymbol{x}-\boldsymbol{u}}{h}\right)\psi_\tau(v)dZ_n^*(v,\boldsymbol{u}), \tag{B.1.30}$$

where $\sigma_{n,*}^2(\boldsymbol{x}) = \mathsf{E}^*\big[\psi_\tau(\varepsilon_i^*)^2\mathbf{1}(|\varepsilon_i^*|<b_n)|\boldsymbol{x}\big]$, and $\Gamma_n^* = \{v : |v| \leq b_n\}$.

$$Y_{1,n}^*(\boldsymbol{x}) = \frac{1}{\sqrt{h^d \hat{f}_{\boldsymbol{X}}(\boldsymbol{x})\sigma_{n,*}^2(\boldsymbol{x})}}\int\int_{\Gamma_n^*} K\left(\frac{\boldsymbol{x}-\boldsymbol{u}}{h}\right)\psi_\tau(v)dB_n^*\big\{\hat{T}(v,\boldsymbol{u})\big\}, \tag{B.1.31}$$

where $B_n^*\big\{\hat{T}(v,\boldsymbol{u})\big\} = W_n^*\big\{\hat{T}(v,\boldsymbol{u})\big\} - \hat{F}(v,\boldsymbol{u})W_n^*(1,...,1)$, W^* is a Brownian motion defined conditional on the sample, and $\hat{T}(v,\boldsymbol{u})$ is the Rosenblatt transformation:

$$\hat{T}(v,\boldsymbol{u}) = \big\{\hat{F}_{X_1|\varepsilon}(u_1|v), \hat{F}_{X_2|\varepsilon}(u_2|u_1,v), ..., \hat{F}_{X_d|X_{d-1},...,X_1,\varepsilon}(u_d|u_{d-1},...,u_1,v), \hat{F}_\varepsilon(v)\big\},$$

given $\hat{F}_{X_1|\varepsilon}(u_1|v), \hat{F}_{X_2|\varepsilon}(u_2|u_1,v), ..., \hat{F}_{X_d|X_{d-1},...,X_1,\varepsilon}(u_d|u_{d-1},...,u_1,v), \hat{F}_\varepsilon(v)$ are associated cdfs obtained from integrating $\hat{f}_{\varepsilon,\boldsymbol{X}}(v,\boldsymbol{u})$.

$$Y_{2,n}^*(\boldsymbol{x}) = \frac{1}{\sqrt{h^d \hat{f}_{\boldsymbol{X}}(\boldsymbol{x})\sigma_{n,*}^2(\boldsymbol{x})}}\int\int_{\Gamma_n^*} K\left(\frac{\boldsymbol{x}-\boldsymbol{u}}{h}\right)\psi_\tau(v)dW_n^*\big\{\hat{T}(v,\boldsymbol{u})\big\}, \tag{B.1.32}$$

$$Y^*_{4,n}(\boldsymbol{x}) = \frac{1}{\sqrt{h^d \hat{f}_{\boldsymbol{X}}(\boldsymbol{x})\sigma^2_{n,*}(\boldsymbol{x})}} \int \sqrt{\hat{f}_{\boldsymbol{X}}(\boldsymbol{u})\sigma^2_{n,*}(\boldsymbol{u})} K\left(\frac{\boldsymbol{x}-\boldsymbol{u}}{h}\right) dW^*_n(\boldsymbol{u}), \quad \text{(B.1.33)}$$

$$Y^*_{5,n}(\boldsymbol{x}) = \frac{1}{\sqrt{h^d}} \int K\left(\frac{\boldsymbol{x}-\boldsymbol{u}}{h}\right) dW^*_n(\boldsymbol{u}). \quad \text{(B.1.34)}$$

From (B.1.30) to (B.1.31) the proof resembles Lemma B.1.3 for quantile regression and B.1.10 for expectile regression. For the bootstrap version of these proofs to hold, it is sufficient to verify the conditions

$$(\log n) h^{-3d} \int_{|v|>c_n} \hat{f}_\varepsilon(v) dv = \mathcal{O}(1), \text{ a.s.} \quad \text{(B.1.35)}$$

for quantile regression and

$$(\log n) h^{-3d} \int_{|v|>c_n} v^2 \hat{f}_\varepsilon(v) dv = \mathcal{O}(1), \text{ a.s.} \quad \text{(B.1.36)}$$

for expectile regression, where $\hat{f}_\varepsilon(v) = (nh_0)^{-1} \sum_{i=1}^n g((v-\hat{\varepsilon}_i)/h_0)$. The rest follows from similar arguments in Lemma B.1.3 and B.1.10.

We will only consider the kernel g with compact support; in particular, with support $[-1,1]$. Via standard arguments one could generalize the result here immediately to, e.g., the Gaussian kernel.

Let $\delta_n = (\log n)^{-1} h^{3d}$. Let $E_n = h^s + (nh^d)^{-1/2} \log n$.

$$\begin{aligned}
\int_{|v|>c_n} \hat{f}_\varepsilon(v) dv &= \frac{1}{nh_0} \sum_{i=1}^n \int_{|v|>c_n} g\left(\frac{\hat{\varepsilon}_i - v}{h_0}\right) dv \\
&\le \frac{1}{nh_0} C_g \sum_{i=1}^n \int_{|v|>c_n} \mathbf{1}(|\hat{\varepsilon}_i - v| \le h_0) dv \\
&\le \frac{1}{nh_0} C_g \sum_{i=1}^n \int_{|v|>c_n} \mathbf{1}(|v| - |\hat{\varepsilon}_i| \le h_0) \mathbf{1}(\hat{\varepsilon}_i - h_0 \le v \le \hat{\varepsilon}_i + h_0) dv \\
&\le \frac{1}{nh_0} C_g \sum_{i=1}^n \mathbf{1}(c_n - h_0 \le |\hat{\varepsilon}_i|) \int_{|v|>c_n} \mathbf{1}(\hat{\varepsilon}_i - h_0 \le v \le \hat{\varepsilon}_i + h_0) dv \\
&\le \frac{2}{n} C_g \sum_{i=1}^n \mathbf{1}(c_n - h_0 \le |\hat{\varepsilon}_i|) \qquad \text{(B.1.37)}
\end{aligned}$$

where C_g is a constant depending on g. For any $\kappa > 0$ and a constant $\lambda > 0$ small

such that $E_n n^\lambda \to 0$ as $n \to \infty$, consider

$$\begin{aligned}
&\mathrm{P}\left\{\left|(1-\mathsf{E})n^{-1}\sum_{i=1}^{n}\mathbf{1}(c_n - h_0 \le |\hat{\varepsilon}_i|)\right| > 2\delta_n\kappa\right\} \le \\
&\mathrm{P}\left\{\left|(1-\mathsf{E})n^{-1}\sum_{i=1}^{n}\mathbf{1}(c_n - h_0 \le |\hat{\varepsilon}_i|)\right| > 2\delta_n\kappa, \left\|\hat{\theta}_n(\boldsymbol{x}) - \theta_0(\boldsymbol{x})\right\| \le E_n n^\lambda\right\} \\
&\qquad\qquad + \mathrm{P}\left\{\left\|\hat{\theta}_n(\boldsymbol{x}) - \theta_0(\boldsymbol{x})\right\| > E_n n^\lambda\right\} \\
&\stackrel{\text{def}}{=} P_{1,n} + P_{2,n}.
\end{aligned}$$

$P_{2,n}$ is summable by similar argument in the proof of Lemma 3.2. Without loss of generality, we assume c_n is large enough so that $h_0 + E_n n^\lambda < c_n/2$ since $h_0, E_n n^\lambda \to 0$. Thus,

$$P_{1,n} \le \mathrm{P}\left\{\left|(1-\mathsf{E})n^{-1}\sum_{i=1}^{n}\mathbf{1}(c_n/2 \le |\varepsilon_i|)\right| > 2\delta_n\kappa\right\}.$$

Let $S_n = \sum_{i=1}^n \mathbf{1}(c_n/2 \le |\varepsilon_i|)$. From (B.1.2) in assumption (C2),

$$\mathrm{Var}(S_n) = n\int_{|v|>c_n/2} f_\varepsilon(v)dv = \mathcal{O}(n^2(\log n)^{-3}h^{6d}) = \mathcal{O}(n^2(\log n)^{-1}\delta_n^2).$$

This yields

$$\begin{aligned}
\sum_{n=1}^{\infty}\mathrm{P}\left(|S_n| > 2\kappa n\delta_n\right) &\le 2\sum_{n=1}^{\infty}\exp\left\{-\frac{4n^2\kappa^2\delta_n^2}{4\,\mathrm{Var}(S_n) + 8n\kappa\delta_n}\right\} \\
&= 2\sum_{n=1}^{\infty}\exp\left\{-\frac{\kappa^2\log n}{1 + 2\kappa\log^2 n/(nh^{3d})}\right\} < \infty, \qquad \text{(B.1.38)}
\end{aligned}$$

given that $\kappa > 1$, since $nh^{3d}(\log n)^{-2} \to \infty$ by assumption (A7). It follows by the Borel-Cantelli lemma that the stochastic part of (B.1.37) is of $\mathcal{O}_p(\delta_n)$. For the expectation, we note that

$$\begin{aligned}
&\mathbf{1}(c_n - h_0 \le |\hat{\varepsilon}_i|) \le \mathbf{1}(c_n - h_0 \le |\varepsilon_i| + \|\hat{\theta}_n(\boldsymbol{x}) - \theta_0(\boldsymbol{x})\|) \\
&\le \mathbf{1}(c_n - h_0 - E_n n^\lambda \le |\varepsilon_i|)\mathbf{1}(\|\hat{\theta}_n(\boldsymbol{x}) - \theta_0(\boldsymbol{x})\| \le E_n n^\lambda) \\
&\qquad\qquad + \mathbf{1}(c_n - h_0 \le |\varepsilon_i| + \|\hat{\theta}_n(\boldsymbol{x}) - \theta_0(\boldsymbol{x})\|)\mathbf{1}(\|\hat{\theta}_n(\boldsymbol{x}) - \theta_0(\boldsymbol{x})\| > E_n n^\lambda) \\
&\le \mathbf{1}(c_n/2 \le |\varepsilon_i|) + \mathbf{1}(\|\hat{\theta}_n(\boldsymbol{x}) - \theta_0(\boldsymbol{x})\| > E_n n^\lambda). \qquad \text{(B.1.39)}
\end{aligned}$$

Therefore,

$$\begin{aligned}
&\mathsf{E}\left[n^{-1}\sum_{i=1}^{n}\mathbf{1}(c_n - h_0 \le |\hat{\varepsilon}_i|)\right] \le \mathsf{E}\left[\mathbf{1}(c_n/2 \le |\varepsilon_i|)\right] + \mathrm{P}\left\{\|\hat{\theta}_n(\boldsymbol{x}) - \theta_0(\boldsymbol{x})\| > E_n n^\lambda\right\} \\
&= \int_{|v|>c_n/2} f_\varepsilon(v)dv + \mathcal{O}(e^{-n^{\mu_1}}) = \mathcal{O}((\log n)^{-3}nh^{6d}),
\end{aligned}$$

for some $\mu_1 > 0$.

Next we show (B.1.36). The sequence c_n will be chosen appropriately later,

$$\begin{aligned}
&\int_{v>c_n} v^2 \hat{f}_\varepsilon(v)dv \\
&\le \frac{1}{nh_0} C_g \sum_{i=1}^n \mathbf{1}(c_n - h_0 \le |\hat{\varepsilon}_i|) \int_{|v|>c_n} v^2 \mathbf{1}(|v| \le h_0 + |\hat{\varepsilon}_i|)dv \\
&\le \frac{1}{nh_0} C_g \sum_{i=1}^n \mathbf{1}(c_n - h_0 \le |\hat{\varepsilon}_i|)(2h_0\hat{\varepsilon}_i^2 + 2h_0^3) \\
&\le \frac{2}{n} C_g \sum_{i=1}^n \hat{\varepsilon}_i^2 \mathbf{1}(c_n - h_0 \le |\hat{\varepsilon}_i|) + \underbrace{\frac{2h_0^2}{n} C_g \sum_{i=1}^n \mathbf{1}(c_n - h_0 \le |\hat{\varepsilon}_i|)}_{T_{1,n}} \\
&\le \frac{4}{n} C_g \sum_{i=1}^n \varepsilon_i^2 \mathbf{1}(c_n - h_0 \le |\hat{\varepsilon}_i|) + \underbrace{\frac{4}{n} C_g \sum_{i=1}^n \left[\hat{\theta}_n(\boldsymbol{X}_i) - \theta_0(\boldsymbol{X}_i)\right]^2 \mathbf{1}(c_n - h_0 \le \hat{\varepsilon}_i)}_{T_{2,n}} + T_{1,n} \\
&= T_{3,n} + T_{2,n} + T_{1,n}. \qquad \text{(B.1.40)}
\end{aligned}$$

Choosing $c_n \asymp (n^{4/b-1}(\log n)^{1+8/b}\delta_n^{-2})^{1/(b-2)}$. Note $c_n > ((\log n)^3(nh^{6d})^{-1})^{1/b}$, and therefore (B.1.2) holds naturally in this case, by assumption (EC1),

$$\int_{|v|>c_n} f_\varepsilon(v)dv \le \int_{|v|>c_n} \frac{|v|^b}{|c_n|^b} f_\varepsilon(v)dv = \mathcal{O}(c_n^b) = \mathcal{O}(n^{4/b-1}(\log n)^{1+8/b}\delta_n^{-2}).$$

It can be shown via similar arguments for showing (B.1.35) that

$$T_{i,n} = \mathcal{O}_p^*((\log n)^{-1}h^{3d}) \text{ a.s. for } i = 1, 2.$$

To bound $T_{3,n}$, given b from (EC1), we choose $M_n = n^{1/b}(\log n)^{2/b}$ and obtain

$$\begin{aligned}
&\mathrm{P}\left\{|(1-\mathsf{E})T_{3,n}| > 2\delta_n\kappa\right\} \\
&\le \mathrm{P}(|(1-\mathsf{E})S_n'| > 2n\kappa\delta_n, \varepsilon_i < M_n, \forall i) + n\mathrm{P}(|\varepsilon_i| \ge M_n) \\
&\qquad + \mathrm{P}\left\{\|\hat{\theta}_n(\boldsymbol{x}) - \theta_0(\boldsymbol{x})\| > E_n n^\lambda\right\} \\
&\stackrel{\text{def}}{=} U_{1,n} + U_{2,n} + U_{3,n},
\end{aligned}$$

where $S_n' = C_g \sum_{i=1}^n \varepsilon_i^2 \mathbf{1}(c_n/2 \le |\varepsilon_i|)$, the term $U_{2,n}$ is of order $o(M_n^{-b})$ by (EC1) and hence summable. $U_{3,n}$ is summable by a similar argument as used in the proof of (B.1.35). Restricting S_n' to the set $\cap_{i=1}^n\{|\varepsilon_i| < M_n\}$, we find

$$\mathrm{Var}(S_n') \le M_n^4 n C_g^2 \int_{c_n/2}^\infty f_\varepsilon(v)dv \le C_{g,b} M_n^4 n c_n^{-b} = \mathcal{O}(n^2(\log n)^{-1}\delta_n^2).$$

This yields

$$\sum_{n=1}^{\infty} U_{1,n} \leq 2 \sum_{n=1}^{\infty} \exp\left\{ -\frac{4n^2\kappa^2\delta_n^2}{4\operatorname{Var}(S_n') + 8n\kappa\delta_n} \right\}$$
$$= 2 \sum_{n=1}^{\infty} \exp\left\{ -\frac{\kappa^2 \log n}{1 + 2\kappa \log^2 n/(nh^{3d})} \right\} < \infty, \tag{B.1.41}$$

given that $\kappa > 1$ and assumption (EA2). It follows by the Borel-Cantelli lemma that $(1 - \mathsf{E})T_{3,n} = \mathcal{O}(\delta_n)$ a.s. It left to control the expectation. By computation in (B.1.39),

$$\mathbf{1}(c_n - h_0 \leq |\hat{\varepsilon}_i|) \leq \mathbf{1}(c_n - h_0 \leq \mathbf{1}(c_n/2 \leq |\varepsilon_i|) + \mathbf{1}(\|\hat{\theta}_n(\boldsymbol{x}) - \theta_0(\boldsymbol{x})\| > E_n n^{\lambda}).$$

Thus, by law of iterative expectation,

$$\mathsf{E}[T_{3,n}] \leq \mathsf{E}\big[\varepsilon_i \mathbf{1}(c_n/2 \leq |\varepsilon_i|)\big] + \mathsf{E}\left[n^{-1} \sum_{i=1}^{n} \varepsilon_i \mathrm{P}\left\{ \|\hat{\theta}_n(\boldsymbol{x}) - \theta_0(\boldsymbol{x})\| > E_n n^{\lambda} \right\} \right]$$
$$= \mathcal{O}(c_n^{2-b}) + \mathcal{O}(e^{-n^{\mu_2}}).$$

It follows immediately by the order of c_n that $\mathsf{E}[T_{3,n}] = \mathcal{O}(\delta_n)$.

In order to show the almost sure uniform convergence of $Y_{4,n}^*(\boldsymbol{x})$ to $Y_{5,n}^*(\boldsymbol{x})$ we need to verify that for quantile regression

$$h^{-d} \log n \sup_{\boldsymbol{x} \in \mathcal{D}} \left| \int_{|v| > c_n} \hat{f}_{\varepsilon|\boldsymbol{X}}(v|\boldsymbol{x}) dv \right| = \mathcal{O}(1), \quad \text{a.s.} \tag{B.1.42}$$

and for expectile regression

$$h^{-d} \log n \sup_{\boldsymbol{x} \in \mathcal{D}} \left| \int_{|v| > c_n} v^2 \hat{f}_{\varepsilon|\boldsymbol{X}}(v|\boldsymbol{x}) dv \right| = \mathcal{O}(1). \quad \text{a.s.} \tag{B.1.43}$$

The first condition can be shown in the same way as showing (B.1.35), and the second one is similar to (B.1.36) given $b \geq 4$. A discretization argument is needed in both cases, but the grid size only grows in polynomial rate in n. The proofs are omitted for brevity.

Using analogous arguments as in Lemma B.1.3 for quantile regression and B.1.10 for expectile regression with (B.1.35) and (B.1.36), it can be shown that $Y_n^*(\boldsymbol{x})$ converges uniformly in probability to $Y_{0,n}^*(\boldsymbol{x})$. The almost sure uniform convergence in probability of $Y_{0,n}^*(\boldsymbol{x})$ to $Y_{5,n}^*(\boldsymbol{x})$ follows by similar arguments in Lemma B.1.4, B.1.5, B.1.8 and B.1.9 for quantile regression and Lemma B.1.11, B.1.12, B.1.14 and B.1.15 for expectile regression, except that $f_{\boldsymbol{X}}(\boldsymbol{x})$, $\sigma_n^2(\boldsymbol{x})$, $F(y, \boldsymbol{x})$ are replaced by $\hat{f}_{\boldsymbol{X}}(\boldsymbol{x})$, $\sigma_{*,n}^2(\boldsymbol{x})$, $\hat{F}(v, \boldsymbol{x})$ respectively, and that the approximation shown in Lemma B.1.6 and B.1.13 is not needed here. Finally, the proof of Theorem 3.1 is completed by an application of the extreme value theorem of Rosenblatt (1976) to $Y_{5,n}^*(\boldsymbol{x})$.

B.2 Supporting lemmas

LEMMA B.2.1 (Kong et al. (2010))**.** Under (A1),(A3)-(A5), for some $s \geq 0$, and $\mathcal{D}$ is an compact subset of $\mathbb{R}^d$. Then

$$\sup_{x\in\mathcal{D}} \left| H_n \left\{ \hat{\beta}(\mathbf{x}) - \beta(\mathbf{x}) \right\} - \beta_n^*(\mathbf{x}) \right| = \mathcal{O}\left(\left\{ \frac{\log n}{nh^d} \right\}^{\lambda(s)} \right). \tag{B.2.1}$$

where

$$\beta_n^*(\mathbf{x}) = -\frac{1}{nh^d} S_{K,g,f}^{-1} H_n^{-1} \left(\sum_{i=1}^{n} K_h(\mathbf{X}_i - \mathbf{x}_i)\varphi(\varepsilon_i) \right) (1, \mathbf{X}_i - \mathbf{x})^\top; \tag{B.2.2}$$

$$\tag{B.2.3}$$

φ is the piecewise derivative of ρ, and

$$\lambda(s) = \min\left\{ \frac{2}{2+s}, \frac{6+2s}{8+4s} \right\}. \tag{B.2.4}$$

Note that under the i.i.d. case, the constant s, which controls the weak dependence, is 0.

LEMMA B.2.2 (Bickel and Wichura (1971): Tightness of processes on a multi-dimensional cube)**.** If $\{X_n\}_{n=1}^{\infty}$ is a sequence in $D[0,1]^d$, $\mathrm{P}(X \in [0,1]^d) = 1$. For neighboring blocks B, C in $[0,1]^d$ (see Definition B.1.1) constants $\lambda_1 + \lambda_2 > 1$, $\gamma_1 + \gamma_2 > 0$, $\{X_n\}_{n=1}^{\infty}$ is tight if

$$\mathsf{E}[|X_n(B)|^{\gamma_1}|X_n(C)|^{\gamma_2}] \leq \mu(B)^{\lambda_1}\mu(C)^{\lambda_2}, \tag{B.2.5}$$

where $\mu(\cdot)$ is a finite nonnegative measure on $[0,1]^d$ (for example, Lebesgue measure), where the increment of X_n on the block B is defined by

$$X_n(B) = \sum_{\boldsymbol{\alpha}\in\{0,1\}^d} (-1)^{d-|\boldsymbol{\alpha}|} X_n\big(\boldsymbol{s} + \boldsymbol{\alpha} \odot (\boldsymbol{t} - \boldsymbol{s})\big).$$

LEMMA B.2.3 (Meerschaert, M. M., Wang, W. and Xiao, Y. (2013))**.** Suppose that $Y = \{Y(\boldsymbol{t}), \boldsymbol{t} \in \mathbb{R}^d\}$ is a centered Gaussian random field with values in $\mathbb{R}$, and denote

$$d(\boldsymbol{s},\boldsymbol{t}) \stackrel{\text{def}}{=} d_Y(\boldsymbol{s},\boldsymbol{t}) = \big(\mathsf{E}|Y(\boldsymbol{t}) - Y(\boldsymbol{s})|^2\big)^{1/2}, \quad \boldsymbol{s},\boldsymbol{t} \in \mathbb{R}^d.$$

Let $\mathcal{D}$ be a compact set contained in a cube with length r in $\mathbb{R}^d$ and let $\sigma^2 = \sup_{\boldsymbol{t}\in\mathcal{D}} \mathsf{E}[Y(\boldsymbol{t})^2]$. For any $m > 0$, $\epsilon > 0$, define

$$\gamma(\epsilon) = \sup_{\boldsymbol{s},\boldsymbol{t}\in\mathcal{D}, \|\boldsymbol{s}-\boldsymbol{t}\|\leq\epsilon} d(\boldsymbol{s},\boldsymbol{t})$$

and

$$Q(m) = (2+\sqrt{2}) \int_1^{\infty} \gamma(m2^{-y^2}) dy.$$

Then for all $a > 0$ which satisfy $a \geq (1 + 4d \log 2)^{1/2}(\sigma + a^{-1})$,

$$\mathrm{P}\left\{\sup_{\boldsymbol{t}\in S} |Y(\boldsymbol{t})| > a\right\} \leq 2^{2d+2}\left(\frac{r}{Q^{-1}(1/a)} + 1\right)^{d} \frac{\sigma + a^{-1}}{a} \exp\left\{-\frac{a^2}{2(\sigma + a^{-1})^2}\right\}, \tag{B.2.6}$$

where $Q^{-1}(a) = \sup\{m : Q(m) \leq a\}$.

Appendix C

Supplementary materials for Chapter 4

C.1 Proof for algorithmic convergence analysis

C.1.1 Proof of Theorem 4.3.2

Proof of Lemma 4.3.1. One can show by elementary matrix algebra that

$$\begin{aligned}\ell(\boldsymbol{\Gamma}, \boldsymbol{\Theta}) &= \sum_{i=1}^{n}\sum_{j=1}^{m} \Theta_{ij}\left(Y_{ij} - \boldsymbol{X}_i^\top \boldsymbol{\Gamma}_{*j}\right) \\ &= \sum_{i=1}^{n}\sum_{j=1}^{m} \Theta_{ij} Y_{ij} - \sum_{i=1}^{n}\sum_{j=1}^{m} \Theta_{ij} \boldsymbol{X}_i^\top \boldsymbol{\Gamma}_{*j} \\ &= \langle \mathbf{Y}, \boldsymbol{\Theta}\rangle + \langle -\mathbf{X}\boldsymbol{\Gamma}, \boldsymbol{\Theta}\rangle.\end{aligned}$$

□

Proof of Theorem 4.3.2. To verify the conditions in Theorem 1 of Nesterov (2005), let $\sigma_2 = 1$, $d(\boldsymbol{\Theta}) = \|\boldsymbol{\Theta}\|_{\mathrm{F}}^2/2$. Lemma 4.3.1 implies that $\hat{\phi}(\boldsymbol{\Theta}) = \langle \mathbf{Y}, \boldsymbol{\Theta}\rangle$, which is convex and continuous. Applying Theorem 1 of Nesterov (2005) yields the desired result. □

C.1.2 Proof of Theorem 4.3.3

Let $\widetilde{L}(\boldsymbol{\Gamma}) = \widehat{Q}_{\tau,\kappa} + \lambda\|\boldsymbol{\Gamma}\|_*$.

$$|L(\boldsymbol{\Gamma}_t) - L(\boldsymbol{\Gamma}^*)| = \left|L(\boldsymbol{\Gamma}_t) - \widetilde{L}(\boldsymbol{\Gamma}_t)\right| + \left|\widetilde{L}(\boldsymbol{\Gamma}_t) - \widetilde{L}(\boldsymbol{\Gamma}^*)\right| + \left|L(\boldsymbol{\Gamma}^*) - \widetilde{L}(\boldsymbol{\Gamma}^*)\right|. \quad \text{(C.1.1)}$$

We have for any $\boldsymbol{\Gamma}$,

$$\widetilde{L}(\boldsymbol{\Gamma}) \le L(\boldsymbol{\Gamma}) \le \widetilde{L}(\boldsymbol{\Gamma}) + \kappa \max_{\boldsymbol{\Theta}\in[\tau-1,\tau]^{n\times m}} \frac{\|W\|_{\mathrm{F}}^2}{2} \le \widetilde{L}(\boldsymbol{\Gamma}) + \kappa\mu(\tau)^2\frac{nm}{2}, \quad \text{(C.1.2)}$$

where the first inequality directly follows from the definition of $f_\kappa(\mathbf{\Gamma}; W)$ in (4.3.3) and the second from the fact that

$$\begin{aligned}\max_{\mathbf{\Theta}\in[\tau-1,\tau]^{n\times m}} \|W\|_{\mathrm{F}}^2 &= \max_{\mathbf{\Theta}\in[\tau-1,\tau]^{n\times m}} \sum_{i\le n, j\le m} \Theta_{ij}^2 \\ &\le \mu(\tau)^2 nm\end{aligned}$$

Hence, for any matrix $\mathbf{\Gamma}$, by the choice of κ in Algorithm 1,

$$\left|L(\mathbf{\Gamma}) - \widetilde{L}(\mathbf{\Gamma})\right| \le \kappa \frac{nm\mu(\tau)^2}{2} \le \frac{\epsilon\mu(\tau)^2}{4}. \tag{C.1.3}$$

Since $\widehat{Q}_{\tau,\kappa}$ is Lipshitz continuous with Lipschitz constant M, by Theorem 4.1 of Ji and Ye (2009) or Theorem 4.4 of Beck and Teboulle (2009) (applied in general real Hilbert space, see their Remark 2.1), we have

$$\left|\widetilde{L}(\mathbf{\Gamma}_t) - \widetilde{L}(\mathbf{\Gamma}^*)\right| \le \frac{2M\|\mathbf{\Gamma}_0 - \mathbf{\Gamma}^*\|_{\mathrm{F}}^2}{(t+1)^2}, \tag{C.1.4}$$

where M is given in Theorem 4.3.2. Combining (C.1.3) and (C.1.4), pick $\kappa = \epsilon/2mn$, insert $M = \frac{2mn}{\epsilon}\|\mathbf{X}\|^2$ by Theorem 4.3.2, (C.1.1) can be estimated by

$$|L(\mathbf{\Gamma}_t) - L(\mathbf{\Gamma}^*)| \le \frac{\epsilon\mu(\tau)^2}{2} + \frac{4mn\|\mathbf{\Gamma}_0 - \mathbf{\Gamma}^*\|_{\mathrm{F}}^2}{(t+1)^2}\frac{\|\mathbf{X}\|^2}{\epsilon}. \tag{C.1.5}$$

Setting the right hand side of (C.1.5) to be ϵ and solve it for T yields the bound (4.3.7).

C.2 Proof of oracle inequalities

C.2.1 Proof of Lemma 4.4.1

The key is the decomposability of the nuclear norm. Again let $\widehat{\mathbf{\Delta}} = \widehat{\mathbf{\Gamma}} - \mathbf{\Gamma}$,

$$\begin{aligned}0 &\le \widehat{Q}_\tau(\mathbf{\Gamma}) - \widehat{Q}_\tau(\widehat{\mathbf{\Gamma}}) \\ &\le \|\nabla\widehat{Q}_\tau(\mathbf{\Gamma})\|\|\widehat{\mathbf{\Delta}}\|_* + \lambda(\|\mathbf{\Gamma}\|_* - \|\widehat{\mathbf{\Gamma}}\|_*) \quad \text{(subgradient condition)} \\ &\le \|\nabla\widehat{Q}_\tau(\mathbf{\Gamma})\|\big(\|\mathcal{P}_{\mathbf{\Gamma}}(\widehat{\mathbf{\Delta}})\|_* + \|\mathcal{P}_{\mathbf{\Gamma}}^\perp(\widehat{\mathbf{\Delta}})\|_*\big) + \lambda(\|\mathcal{P}_{\mathbf{\Gamma}}(\mathbf{\Gamma})\|_* - \|\mathcal{P}_{\mathbf{\Gamma}}^\perp(\widehat{\mathbf{\Gamma}})\|_* - \mathcal{P}_{\mathbf{\Gamma}}(\widehat{\mathbf{\Gamma}})) \\ &\qquad\qquad \text{(Decomposability } \|\cdot\|_*) \\ &\le \|\nabla\widehat{Q}_\tau(\mathbf{\Gamma})\|\big(\|\mathcal{P}_{\mathbf{\Gamma}}(\widehat{\mathbf{\Delta}})\|_* + \|\mathcal{P}_{\mathbf{\Gamma}}^\perp(\widehat{\mathbf{\Delta}})\|_*\big) + \lambda(\|\mathcal{P}_{\mathbf{\Gamma}}(\widehat{\mathbf{\Delta}})\|_* - \|\mathcal{P}_{\mathbf{\Gamma}}^\perp(\widehat{\mathbf{\Delta}})\|_*).\end{aligned}$$

Rearrange to get,

$$(\lambda - \|\nabla\widehat{Q}_\tau(\mathbf{\Gamma})\|)\|\mathcal{P}_{\mathbf{\Gamma}}^\perp(\widehat{\mathbf{\Delta}})\|_* \le (\lambda + \|\nabla\widehat{Q}_\tau(\mathbf{\Gamma})\|)\|\mathcal{P}_{\mathbf{\Gamma}}(\widehat{\mathbf{\Delta}})\|_*.$$

Choose $\lambda \ge 2\|\nabla\widehat{Q}_\tau(\mathbf{\Gamma})\|$,

$$\frac{1}{2}\lambda\|\mathcal{P}_{\mathbf{\Gamma}}^\perp(\widehat{\mathbf{\Delta}})\|_* \le \frac{3}{2}\lambda\|\mathcal{P}_{\mathbf{\Gamma}}(\widehat{\mathbf{\Delta}})\|_*.$$

Hence, $\|\mathcal{P}_{\mathbf{\Gamma}}^\perp(\widehat{\mathbf{\Delta}})\|_* \le 3\|\mathcal{P}_{\mathbf{\Gamma}}(\widehat{\mathbf{\Delta}})\|_*$. Note that this condition appears also in Negahban and Wainwright (2011) Eqn. (12) in pp. 1077.

C.2.2 Proof of Lemma 4.4.2

1. Let $Q_{\tau,j}(\mathbf{\Gamma}_{*j}) = \mathsf{E}[\rho_\tau(Y_{ij} - \boldsymbol{X}_i^\top \mathbf{\Gamma}_{*j})]$. By Assumption 4.3,

$$\begin{aligned}
&Q_{\tau,j}(\mathbf{\Gamma}_{*j} + \mathbf{\Delta}_{*j}) - Q_{\tau,j}(\mathbf{\Gamma}_{*j}) \\
&= \mathsf{E}\bigg[\int_0^{\boldsymbol{X}_i^\top \mathbf{\Delta}_{*j}} F_{Y_j|\boldsymbol{X}_i}(\boldsymbol{X}_i^\top \mathbf{\Gamma}_{*j} + z) - F_{Y_j|\boldsymbol{X}_i}(\boldsymbol{X}_i^\top \mathbf{\Gamma}_{*j}) dz\bigg] \\
&= \mathsf{E}\bigg[\int_0^{\boldsymbol{X}_i^\top \mathbf{\Delta}_{*j}} z f_{Y_j|\boldsymbol{X}_i}(\boldsymbol{X}_i^\top \mathbf{\Gamma}_{*j}) + \frac{z^2}{2} f'_{Y_j|\boldsymbol{X}_i}(\boldsymbol{X}_i^\top \mathbf{\Gamma}_{*j} + z^\dagger) dz\bigg] \\
&\geq \underline{f} \frac{\mathsf{E}\big[(\boldsymbol{X}_i^\top \mathbf{\Delta}_{*j})^2\big]}{4} + \underline{f} \frac{\mathsf{E}\big[(\boldsymbol{X}_i^\top \mathbf{\Delta}_{*j})^2\big]}{4} - \frac{1}{6}\bar{f}' \mathsf{E}[|\boldsymbol{X}_i^\top \mathbf{\Delta}_{*j}|^3]
\end{aligned}$$

   for $z^\dagger \in [0, z]$. Notice the condition that $\|\mathbf{\Delta}\|_{L_2(\Pi)} \leq 4\nu$ implies

$$\underline{f} \frac{\mathsf{E}\big[(\boldsymbol{X}_i^\top \mathbf{\Delta}_{*j})^2\big]}{4} > \frac{1}{6}\bar{f}' \mathsf{E}[|\boldsymbol{X}_i^\top \mathbf{\Delta}_{*j}|^3]$$

   Therefore,

$$Q_\tau(\mathbf{\Gamma} + \mathbf{\Delta}) - Q_\tau(\mathbf{\Gamma}) \geq \underline{f} m^{-1} \sum_{j=1}^m \frac{\mathsf{E}(\boldsymbol{X}_i^\top \mathbf{\Delta}_{*j})^2}{4} = \frac{1}{4}\underline{f} \|\mathbf{\Delta}\|_{L_2(\Pi)}^2.$$

2. By the decomposability of nuclear norm, $\mathbf{\Delta} \in \mathcal{K}(\mathbf{\Gamma}, 3)$ and Assumption 4.4, we can estimate

$$\begin{aligned}
\|\mathbf{\Delta}\|_* &= \|\mathcal{P}_{\mathbf{\Gamma}}(\mathbf{\Delta})\|_* + \|\mathcal{P}_{\mathbf{\Gamma}}^\perp(\mathbf{\Delta})\|_* \leq 4\|\mathcal{P}_{\mathbf{\Gamma}}(\mathbf{\Delta})\|_* \leq 4\sqrt{2r}\|\mathcal{P}_{\mathbf{\Gamma}}(\mathbf{\Delta})\|_{\mathrm{F}} \\
&\leq \frac{4\sqrt{2r}}{\beta_{\mathbf{\Gamma},3}} \|\mathbf{\Delta}\|_{L_2(\Pi)}.
\end{aligned}$$

C.2.3 Proof of Lemma 4.4.3

We restrict on the event Ω that assumption 4.2 holds. Observe that $\rho_\tau\big\{Y_{ij} - \boldsymbol{X}_i(\mathbf{\Gamma}_{*j} + \mathbf{\Delta}_{*j})\big\} - \rho_\tau\big\{Y_{ij} - \boldsymbol{X}_i \mathbf{\Gamma}_{*j}\big\}$ boils down to

$$\begin{aligned}
&\rho_\tau\big\{Y_{ij} - \boldsymbol{X}_i^\top(\mathbf{\Gamma}_{*j} + \mathbf{\Delta}_{*j})\big\} - \rho_\tau\big\{Y_{ij} - \boldsymbol{X}_i^\top \mathbf{\Gamma}_{*j}\big\} \\
=&\big\{\tau - \mathbf{1}(\varepsilon_{ij} \leq \boldsymbol{X}_i^\top \mathbf{\Delta}_{*j})\big\}(\varepsilon_{ij} - \boldsymbol{X}_i^\top \mathbf{\Delta}_{*j}) - \big\{\tau - \mathbf{1}(\varepsilon_{ij} \leq 0)\big\}\varepsilon_{ij} \\
=&\varepsilon_{ij}\mathbf{1}(\boldsymbol{X}_i^\top \mathbf{\Delta}_{*j} < \varepsilon_{ij} \leq 0) - \varepsilon_{ij}\mathbf{1}(0 < \varepsilon_{ij} \leq \boldsymbol{X}_i^\top \mathbf{\Delta}_{*j}) \\
&\qquad - \boldsymbol{X}_i^\top \mathbf{\Delta}_{*j}\big\{\tau - \mathbf{1}(\varepsilon_{ij} \leq \boldsymbol{X}_i^\top \mathbf{\Delta}_{*j})\big\}, \qquad \text{(C.2.1)}
\end{aligned}$$

where $\tau - \mathbf{1}(\varepsilon_{ij} \leq \boldsymbol{X}_i^\top \mathbf{\Delta}_{*j})$ and $\varepsilon_{ij} = Y_{ij} - \boldsymbol{X}_i^\top \mathbf{\Gamma}_{*j}$ are independent across i and j by Assumption 4.1. By triangle inequality and (C.2.1), $\mathcal{A}(t) \leq \mathcal{B}(t) + \mathcal{C}(t) + \mathcal{D}(t)$ where

$$\mathcal{B}(t) = \sup_{\|\mathbf{\Delta}\|_{L_2(\Pi)} \leq t, \mathbf{\Delta} \in \mathcal{K}(\mathbf{\Gamma}, 3)} \bigg| \mathbb{G}_n \bigg[\frac{1}{m} \sum_{j=1}^m \varepsilon_{ij} \mathbf{1}(\boldsymbol{X}_i^\top \mathbf{\Delta}_{*j} < \varepsilon_{ij} \leq 0)\bigg]\bigg|,$$

$$\mathcal{C}(t) = \sup_{\|\boldsymbol{\Delta}\|_{L_2(\Pi)} \le t, \boldsymbol{\Delta} \in \mathcal{K}(\boldsymbol{\Gamma}, 3)} \left| \mathbb{G}_n \left[\frac{1}{m} \sum_{j=1}^{m} \varepsilon_{ij} \mathbf{1}(0 < \varepsilon_{ij} \le \boldsymbol{X}_i^\top \boldsymbol{\Delta}_{*j}) \right] \right|,$$

$$\mathcal{D}(t) = \sup_{\substack{\|\boldsymbol{\Delta}\|_{L_2(\Pi)} \le t, \\ \boldsymbol{\Delta} \in \mathcal{K}(\boldsymbol{\Gamma}, 3)}} \left| \mathbb{G}_n \left[\frac{1}{m} \sum_{j=1}^{m} \boldsymbol{X}_i^\top \boldsymbol{\Delta}_{*j} \{ \tau - \mathbf{1}(\varepsilon_{ij} \le \boldsymbol{X}_i^\top \boldsymbol{\Delta}_{*j}) \} \right] \right|.$$

First, we consider $\mathcal{B}(t)$. Condition on $\boldsymbol{X}$, the variable $\varepsilon_{ij} \mathbf{1}(\boldsymbol{X}_i^\top \boldsymbol{\Delta}_{*j} < \varepsilon_{ij} \le 0)$ lies in $(\boldsymbol{X}_i^\top \boldsymbol{\Delta}_{*j}, 0]$. Condition on $\boldsymbol{X}_i$, applying Hoeffding's inequality C.3.4 gives

$$\begin{aligned} &\mathrm{P}\left(\left| \mathbb{G}_n \left[\frac{1}{m} \sum_{j=1}^{m} \varepsilon_{ij} \mathbf{1}(\boldsymbol{X}_i^\top \boldsymbol{\Delta}_{*j} < \varepsilon_{ij} \le 0) \right] \right| \ge s \right) \\ &\qquad \le 2 \exp\left(- \frac{2m^2 s^2}{n^{-1} \sum_{i=1}^{n} \sum_{j=1}^{m} (\boldsymbol{X}_i^\top \boldsymbol{\Delta}_{*j})^2} \right) \\ &\qquad \le 2 \exp\left(- \frac{2m^2 s^2}{\alpha^2 \|\boldsymbol{\Delta}\|_{L_2(\Pi)}^2 c_2 \sigma_{\max}(\boldsymbol{\Sigma}_{\boldsymbol{X}})} \right), \end{aligned}$$

where the second inequality comes from Assumption 4.2, Lemma 4.4.2 (ii) and Hölder's inequality; more explicitly,

$$\begin{aligned} n^{-1} \sum_{i=1}^{n} \sum_{j=1}^{m} (\boldsymbol{X}_i^\top \boldsymbol{\Delta}_{*j})^2 &\le n^{-1} \operatorname{tr}(\boldsymbol{\Delta} \boldsymbol{X}^\top \boldsymbol{X} \boldsymbol{\Delta}) \le \|\boldsymbol{\Delta}\|_*^2 \left\| n^{-1} \mathbf{X}^\top \mathbf{X} \right\| \\ &\le \alpha^2 c_2 \|\boldsymbol{\Delta}\|_{L_2(\Pi)}^2 \|\boldsymbol{\Sigma}_{\boldsymbol{X}}\|. \end{aligned} \tag{C.2.2}$$

Hence, $\mathcal{B}(t) \le \frac{\alpha t \sqrt{c_2 \|\boldsymbol{\Sigma}_{\boldsymbol{X}}\| \log(p+m)}}{m}$ with probability greater than $1 - 2(p+m)^{-2}$.

Performing similar procedure to $\mathcal{C}(t)$ as for bounding $\mathcal{B}(t)$, we also have

$$\mathcal{C}(t) \le \frac{\alpha t \sqrt{c_2 \|\boldsymbol{\Sigma}_{\boldsymbol{X}}\| \log(p+m)}}{m} \text{ with probability } \ge 1 - 2(p+m)^{-2}.$$

Next we consider $\mathcal{D}(t)$. Condition on $\boldsymbol{X}_i$, $\tau - \mathbf{1}(\varepsilon_{ij} \le \boldsymbol{X}_i^\top \boldsymbol{\Delta}_{*j})$ is independent across i and j and is bounded by $\tau \vee (1-\tau)$. Using Hoeffding's inequality C.3.3,

$$\begin{aligned} &\mathrm{P}\left(\left| \mathbb{G}_n \left[\frac{1}{m} \sum_{j=1}^{m} \boldsymbol{X}_i^\top \boldsymbol{\Delta}_{*j} \{ \tau - \mathbf{1}(\varepsilon_{ij} \le \boldsymbol{X}_i^\top \boldsymbol{\Delta}_{*j}) \} \right] \right| \ge s \right) \\ &\qquad \le \exp\left(1 - \frac{C' s^2 m^2}{\{\tau \vee (1-\tau)\} n^{-1} \sum_{i=1}^{n} \sum_{j=1}^{m} (\boldsymbol{X}_i^\top \boldsymbol{\Delta}_{*j})^2} \right) \\ &\qquad \le \exp\left(1 - \frac{C' s^2 m^2}{\{\tau \vee (1-\tau)\} \alpha^2 \|\boldsymbol{\Delta}\|_{L_2(\Pi)}^2 c_2 \sigma_{\max}(\boldsymbol{\Sigma}_{\boldsymbol{X}})} \right), \end{aligned}$$

where the second inequality follows from the same deduction in (C.2.2). Therefore,

$$\begin{aligned} \mathcal{D}(t) \le & \frac{\alpha t \sqrt{2\{\tau \vee (1-\tau)\} c_2 \|\boldsymbol{\Sigma}_{\boldsymbol{X}}\| \log(p+m)}}{\sqrt{C'} m} \\ & \text{with probability greater than } 1 - 3(p+m)^{-2}. \end{aligned}$$

Summing up the results for $\mathcal{B}(t)$, $\mathcal{C}(t)$ and $\mathcal{D}(t)$, with the restriction on the event Ω, we obtain

$$\mathcal{A}(t) \leq \left(\sqrt{\frac{2\tau \vee (1-\tau)}{C'}} + 2\right) \frac{\alpha t \sqrt{c_2 \|\boldsymbol{\Sigma}_{\boldsymbol{X}}\| \log(p+m)}}{m}$$
$$\text{with probability greater than } 1 - 9(p+m)^{-2} - \gamma_n,$$

as $e < 3$

C.2.4 Proof of Lemma 4.4.5

Applying the same $\mathcal{E}$-net argument on the unit m dimensional Euclidean sphere $S^{m-1} = \{\mathbf{u} \in \mathbb{R}^m : \|\mathbf{u}\|_2 = 1\}$ as in the proof of Lemma 3 in Negahban and Wainwright (2011), we obtain

$$\begin{aligned} \mathrm{P}\left(\frac{1}{n}\|\mathbf{X}^\top \mathbf{W}\| \geq 4s\right) &= \mathrm{P}\left(\sup_{\substack{\mathbf{v}\in S^{p-1} \\ \mathbf{u}\in S^{m-1}}} \frac{1}{n}\left|\mathbf{v}^\top \mathbf{X}^\top \mathbf{W}\mathbf{u}\right| \geq 4s\right) \\ &\leq 8^{p+m} \sup_{\substack{\mathbf{v}\in S^{p-1}, \mathbf{u}\in S^{m-1} \\ \|\mathbf{u}\|=\|\mathbf{v}\|=1}} \mathrm{P}\left(\frac{|\langle \mathbf{X}\mathbf{v}, \mathbf{W}\mathbf{u}\rangle|}{n} \geq s\right). \end{aligned} \tag{C.2.3}$$

To bound $n^{-1}\langle \mathbf{X}\mathbf{v}, \mathbf{W}\mathbf{u}\rangle = n^{-1}\sum_{i=1}^n \langle \mathbf{v}, \boldsymbol{X}_i\rangle\langle \mathbf{u}, \boldsymbol{W}_i\rangle$, first we show the sub-Gaussianity of $\langle \mathbf{u}, \boldsymbol{W}_i\rangle$. Since $|W_{ij}| \leq \tau \vee (1-\tau)$. It follows by Hoeffding's inequality C.3.3 that

$$\mathrm{P}\big(\langle \mathbf{u}, \boldsymbol{W}_i\rangle \geq s\big) \leq \exp\left(1 - \frac{C's^2}{\{\tau \vee (1-\tau)\}\|\mathbf{u}\|_2^2}\right) = \exp\left(1 - \frac{C's^2}{\tau \vee (1-\tau)}\right).$$

It can also be concluded that (see Definition 5.7 and discussion of Vershynin (2012)) $\|\langle \mathbf{u}, \boldsymbol{W}_i\rangle\|_{\psi_2} = \sqrt{\tau \vee (1-\tau)}$.

We apply Hoeffding's inequality C.3.3 again to bound $n^{-1}\sum_{i=1}^n \langle \mathbf{v}, \boldsymbol{X}_i\rangle\langle \mathbf{u}, \boldsymbol{W}_i\rangle$. Conditioning on $\boldsymbol{X}_i$, we have

$$\begin{aligned} \mathrm{P}\left(\left|n^{-1}\sum_{i=1}^n \langle \mathbf{v}, \boldsymbol{X}_i\rangle\langle \mathbf{u}, \boldsymbol{W}_i\rangle\right| \geq s\right) &\leq \exp\left(1 - \frac{C'ns^2}{\{\tau \vee (1-\tau)\}n^{-1}\sum_{i=1}^n \langle \mathbf{v}, \boldsymbol{X}_i\rangle^2}\right) \\ &\leq \exp\left(1 - \frac{C'ns^2}{\{\tau \vee (1-\tau)\}c_2\|\boldsymbol{\Sigma}_{\boldsymbol{X}}\|}\right). \end{aligned}$$

where the second inequality follows from the fact that $\|\mathbf{v}\|_2 = 1$ and $n^{-1}\sum_{i=1}^n \langle \mathbf{v}, \boldsymbol{X}_i\rangle^2 \leq \|\mathbf{X}^\top \mathbf{X}/n\| \leq c_2\|\boldsymbol{\Sigma}_{\boldsymbol{X}}\|$ on the event that Assumption 4.2 holds.

To summarize, on the event that Assumption 4.2 holds,

$$\begin{aligned} \mathrm{P}\left(\frac{1}{n}\|\mathbf{X}^\top \mathbf{W}\| \geq 4s\right) &\leq 8^{p+m} \exp\left(1 - \frac{C'ns^2}{\{\tau \vee (1-\tau)\}c_2\|\boldsymbol{\Sigma}_{\boldsymbol{X}}\|}\right) \\ &\leq \exp\left(1 - \frac{C'ns^2}{\{\tau \vee (1-\tau)\}c_2\|\boldsymbol{\Sigma}_{\boldsymbol{X}}\|} + (p+m)\log 8\right). \end{aligned}$$

Therefore,

$$\frac{1}{n}\|\mathbf{X}^\top\mathbf{W}\| \leq 4\cdot\sqrt{2\log 8\frac{\{\tau\vee(1-\tau)\}c_2\|\mathbf{\Sigma}_{\boldsymbol{X}}\|}{C'}}\sqrt{\frac{p+m}{n}},$$

with probability greater than $1-3e^{-(p+m)\log 8}-\gamma_n$, as $e<3$.

C.3 Supplementary lemmas

Definition C.3.1. Let $\mathcal{X}=\mathbb{R}^{p\times n}$ with inner product $\langle\mathbf{A},\mathbf{B}\rangle=\mathrm{tr}(\mathbf{A}^\top\mathbf{B})$ and $\|\cdot\|$ be the induced norm. $f:\mathcal{X}\to\mathbb{R}$ a lower semicontinuous convex function. The *proximity operator of* f, $S_f:\mathcal{X}\to\mathcal{X}$:

$$S_f(\mathbf{Y})\stackrel{\text{def}}{=}\arg\min_{\mathbf{X}\in\mathcal{X}}\left\{f(\mathbf{X})+\frac{1}{2}\|\mathbf{X}-\mathbf{Y}\|^2\right\},\forall\mathbf{Y}\in\mathcal{X}.$$

THEOREM C.3.2 (Theorem 2.1 of Cai et al. (2010))**.** Suppose the singular decomposition of $\mathbf{Y}=\mathbf{UDV}^\top\in\mathbb{R}^{p\times m}$, where $\mathbf{D}$ is a $p\times m$ rectangular diagonal matrix and $\mathbf{U}$ and $\mathbf{V}$ are unitary matrices. The proximity operator $S_\lambda(\cdot)$ associated with $\lambda\|\cdot\|_*$ is

$$S_\lambda(\mathbf{Y})\stackrel{\text{def}}{=}\mathbf{U}(\mathbf{D}-\lambda\mathbf{I}_{pm})_+\mathbf{V}^\top, \tag{C.3.1}$$

where $\mathbf{I}_{pm}$ is the $p\times m$ rectangular identity matrix with diagonal elements equal to 1.

LEMMA C.3.3 (Hoeffding's Inequality, Proposition 5.10 of Vershynin (2012))**.** Let $X_1,...,X_n$ be independent centered sub-gaussian random variables, and let $K=\max_i\|X_i\|_{\psi_2}$. Then for every $\mathbf{a}=(a_1,...,a_n)\in\mathbb{R}^n$ and every $t\geq 0$, we have

$$\mathrm{P}\left(\left|\sum_{i=1}^n a_iX_i\right|\geq t\right)\leq e\cdot\exp\left(-\frac{C't^2}{K^2\|\mathbf{a}\|_2^2}\right),$$

where $C'>0$ is a universal constant.

LEMMA C.3.4 (Hoeffding's Inequality: classical form)**.** Let $X_1,...,X_n$ be independent random variables such that $X_i\in[a_i,b_i]$ almost surely, then

$$\mathrm{P}\left(\left|\sum_{i=1}^n X_i\right|\geq t\right)\leq 2\exp\left(-\frac{2n^2t^2}{\sum_{i=1}^n(b_i-a_i)^2}\right).$$

LEMMA C.3.5 (Wainwright (2009))**.** Let $\mathbf{X}\in\mathbb{R}^{n\times p}$ be a random matrix with i.i.d. rows sampled from a p-variate $N(0,\mathbf{\Sigma}_{\boldsymbol{X}})$ distribution. Then for $n\geq 2m$, we have

$$\mathrm{P}\left[\sigma_{\min}\left(\frac{1}{n}\mathbf{X}^\top\mathbf{X}\right)\geq\frac{\sigma_{\min}(\mathbf{\Sigma}_{\boldsymbol{X}})}{9},\sigma_{\max}\left(\frac{1}{n}\mathbf{X}^\top\mathbf{X}\right)\leq 9\sigma_{\max}(\mathbf{\Sigma}_{\boldsymbol{X}})\right]\geq 1-4\exp(-n/2).$$

One may see the discussion after the proof of Lemma 9 in Wainwright (2009) for details.

Selbständigkeitserklärung

Ich bezeuge durch meine Unterschrift, dass meine Angaben über die bei der Abfassung meiner Dissertation benutzten Hilfsmittel, über die mir zuteil gewordene Hilfe sowie über frühere Begutachtungen meiner Dissertation in jeder Hinsicht der Wahrheit entsprechen.

Berlin, 13 April 2015

Shih-Kang Chao